Masakhane Playbook

Democratizing machine translation for African languages

Table of contents:

- Key Components of Planning
  - Budgeting Annotation Costs
  - Time Estimation
  - Scaling Strategy

Data Cleaning and Preprocessing

- Why Data Cleaning Matters
- Key Steps in Data Cleaning
  - Deduplication, Normalization, and Filtering
  - Language Detection and Formatting
  - Noise and Toxicity Handling
  - Missing and Corrupted Data Handling

Data Provenance and Traceability

- Why Provenance Matters
- Key Components of Provenance
  - Source Tracking
  - Data Lineage
  - Transformation Logs

Ethics, Bias, and Governance

- Why Ethics and Governance Matter
- Key Components of Ethical Data Practices
  - Bias Identification and Mitigation
  - PII Detection and Removal
  - Anonymization Strategies
  - Sensitive Attribute and Content Handling
  - Fair Representation
  - Risk Documentation and Transparency
-  Annotation Task Design and Human Factors
-  Inclusive and Bias-Aware Annotation
-  Training and Guidelines
-  Workflow and Adjudication

Annotation Task Design and Human Factors

- Task complexity estimation
- Annotator fatigue
- UI/UX considerations

Inclusive and Bias-Aware Annotation

- Why Inclusion and Bias Awareness Matter
- Key Components
  - Gender, Age, and Cultural Diversity
  - Recording Annotator Personas

- Bias Awareness Training
- Community Participation

Training and Guidelines

- Annotation Guidelines with Examples
- Training and Calibration Rounds
- Pilot Annotation and Iteration
- Minimum Viable Dataset per Task

Workflow and Adjudication

- Multi-annotator setup
- Task assignment and redundancy
- Disagreement resolution and expert adjudication
- Data Quality Management
- Data Quality Assurance

Data Quality Management

- Class imbalance handling
- Outlier and noise detection
- Error analysis pipelines
- Dataset versioning and updates

Data Quality Assurance

- Inter-Annotator Agreement (Kappa, Alpha)
- Gold data and control checks
- Auditing and spot-checking
- Feedback loops
- Image Data
- Multimodal Data
- Speech Data
- Text Data

Image Data

Multimodal Data

Speech Data

Text Data

- Templates & Artifacts
- Consent Form Template
- Contributor Agreement Template
- Data Ownership Documentation Template
- Licensing and Compliance
- Sustainability Plan Template
- Documentation and Reporting
- Tooling and Infrastructure

- Annotation quality verified
- Documentation completed
- Licensing defined
- Ethical review conducted
- Baselines and splits provided
- Public access ensured
- Collaboration and Shared Tasks
- Community Ecosystems

Collaboration and Shared Tasks

- Shared tasks and benchmarks
- Workshops and open challenges

Community Ecosystems

- Community initiatives (Masakhane, EthioNLP, HausaNLP)
- Academic and industry collaboration
- Contribution and contributor guidelines
- Defining text classification tasks:
- Data sources
- Data Collection and Selection Approaches
- Data Processing and Sampling
- Annotation Tools
- Annotator Recruitment/Selection
- Annotation Quality Control
- Annotation Agreement
- Sentiment Analysis
- Emotion Analysis
- Hate Speech Analysis
- Data Quality Control
- Annotator Safety and Mental Health

Defining text classification tasks:

Data sources

Data Collection and Selection Approaches

Data Processing and Sampling

Annotation Tools

-
- 5.1. Crowdsourcing Platforms
- 5.2. In-House (Self-Hosted) Tools
- 5.3. Lightweight Tools for Small Datasets

Annotator Recruitment/Selection

Annotation Quality Control

1. Introduction

A comprehensive guide to dataset design, annotation, and task formulation for building reliable and responsible language AI systems.

> 📚 **CITING THIS PLAYBOOK**
>
> Using this resource in research, teaching, or a project? **Jump to the citation block** at the bottom of this page for BibTeX, APA, and other formats. The full citation page lives at `/cite`.

> 🤝 **HELP BUILD THE PLAYBOOK**
>
> This is a **community-driven** resource. If you spot a gap, want to write a chapter, translate a page, or suggest an improvement — contributions from researchers, practitioners, students, and language experts are very welcome. See the **contribution guide** to get started, or join the conversation on Discord.

Welcome to the dataset design and annotation playbook!

This playbook will help you plan and develop **training and evaluation datasets**, define **annotation schemas**, and design **AI tasks** across different languages, domains, and modalities. It provides guidance on dataset structuring, labeling strategies, and ethical considerations for language technologies.

How to read this playbook

The playbook is organised end-to-end through the dataset lifecycle, but you don't have to read it linearly. Pick the path that fits where you are:

- **New to dataset design.** Start here, then read **chapters 2–4** in order — Data Collection → Annotation Design → Data Quality. They build on each other and cover the foundations everyone needs.

- **You already have raw data, want guidance on annotation.** Jump to **chapter 3 (Annotation Design and Workforce Management)**, then **chapter 4 (Data Quality Assurance and Validation)**.
- **You're working with a specific modality** (speech, multimodal, low-resource scripts). Skip to **chapter 5 (Modality-Specific Task Design)**.
- **You're using LLMs to generate or augment data.** Read **chapter 7 (LLM-Assisted and Synthetic Data Generation)** for the trade-offs and safeguards.
- **You're preparing a dataset for release or publication.** Read **chapter 6 (Documentation, Data Release, and Governance)** and **chapter 9 (Dataset Lifecycle Management and Release Checklist)**.
- **You're a coordinator onboarding a team or community group.** See Onboarding a Team and Running a Playbook Workshop.
- **You're reading offline or on a slow connection.** Use **Download PDF** in the navbar — the entire playbook bundles into a single file, regenerated automatically on every release.
- **You'd rather read in Hausa, Amharic, Swahili, French, or Portuguese.** Use the language switcher in the top-right of the navbar. Translations are community-maintained and grow over time.

Throughout the playbook, you'll find practical templates (consent forms, annotation guidelines, governance checklists), worked examples from real African-language projects, and links to source datasets and tools you can reuse.

Who is this playbook for?

This playbook is designed for:

- **Researchers** working on NLP dataset creation and evaluation
- **Annotation teams** developing labeled datasets
- **Project managers and coordinators** overseeing data collection and annotation workflows
- **AI practitioners** designing and evaluating language models
- **Students and academics** studying dataset design and annotation
- **Multilingual communities** contributing to language resources

- **Trainers and facilitators** who run workshops or onboarding sessions for contributors

What will you learn?

By the end of this playbook, you will understand:

- How to define the **purpose and scope** of a dataset
- Differences between **training and evaluation datasets**
- Trade-offs between **scale and quality**
- How to design **label schemas and ontologies**
- Approaches for **multi-label, single-label, and structured outputs**
- How to handle **ambiguity, edge cases, and annotation boundaries**
- Best practices for **multilingual and cross-lingual dataset design**
- Ethical considerations, risks, and limitations in dataset creation

How to use this playbook

Each section of this playbook contains:

- **Clear explanations** of dataset design principles
- **Structured guidance** for task and schema definition
- **Examples and edge cases** to support annotation decisions
- **Practical recommendations** for dataset creation workflows
- **Ethical considerations** to guide responsible use

Getting Started

Ready to begin? Start with our foundational sections:

1. **Purpose of this Playbook** – Understand target users, scope, and intended use
2. **How to Use This Playbook** – Learn how to navigate chapters and contribute
3. **Dataset Types and Design Goals** – Explore dataset categories and trade-offs
4. **Task and Schema Definition** – Define tasks, labels, and annotation structures

5. **Glossary and Terminology** – Learn key concepts and definitions

Purpose of this playbook

- Target users and communities
- Languages, domains, and modalities covered
- Intended use and risks

Dataset Types and Design Goals

- Training vs evaluation datasets
- General-purpose vs domain-specific datasets
- Scale vs quality trade-offs
- Monolingual, multilingual, cross-lingual setups

Task and Schema Definition

- Task formulation (classification, generation, alignment, retrieval)
- Label schema and ontology design
- Multi-label vs single-label vs structured outputs
- Ambiguity, edge cases, and annotation boundaries

Glossary and Terminology

A reference section providing clear definitions of the key terms used throughout the playbook — see the **Glossary** for definitions of *annotation, inter-annotator agreement, Cohen's kappa, low-resource language, modality,* and other terms.

How to cite this playbook

If the AfriPlaybook informs your research, teaching, or project, please cite it.

BibTeX:

```
@misc{masakhane2026playbook,
  author       = {{Masakhane Community}},
  title        = {AfriPlaybook: A Practical Guide for Building NLP Systems
for African Languages},
  year         = {2026},
  publisher    = {Masakhane},
  url          = {https://warakacommunity.github.io/AfriPlaybook/},
  note         = {Open-source community resource}
}
```

Plain text (APA-style):

> Masakhane Community. (2026). *AfriPlaybook: A Practical Guide for Building NLP Systems for African Languages*. https://warakacommunity.github.io/AfriPlaybook/

For other formats (MLA, Chicago, etc.) and a machine-readable `CITATION.cff`, see the /cite page.

If you reference a specific chapter, please include the chapter title and its URL.

Cite this page 4 min read

*Last updated on **Jun 15, 2026** by **Seid Muhie Yimam***

Onboarding a Team

This page is for coordinators and project leads who want to introduce the playbook to a new team or community group.

Where to Start by Role

Rather than sharing the full playbook with a new team, start with the chapter most relevant to their immediate task.

Role	Recommended starting point
Annotator	Training and Guidelines, then the relevant modality chapter
Voice recorder	Speech Data
Reviewer / quality checker	Data Quality Assurance and Validation
Coordinator	Cost and Resource Planning → Annotation Design
Linguist	Annotation Design → Inclusive and Bias-Aware Design
Dataset release lead	Documentation and Governance → Dataset Lifecycle

First-Session Structure

A single 60–90 minute orientation session is usually enough to get a team started:

1. Walk through the relevant chapter together (screen share or printed copy)
2. Work through one concrete example from the actual project — not an abstract sample
3. Run a short practice task and discuss any questions before independent work begins

Onboarding Materials to Prepare

- [ ] Link (or printout) of the relevant chapter
- [ ] 3–5 real annotation or recording examples from your project
- [ ] Contact details for who to reach with tool or task questions
- [ ] Offline copy of materials for contributors without reliable internet (see Inclusive Access)

Checking Readiness

Before a contributor starts independent work, confirm they can:

- Navigate to the relevant playbook section without help
- Correctly identify the label or action for at least three practice examples
- Reach their point of contact if something is unclear

Contributors who cannot pass a short readiness check benefit from a second orientation rather than starting at full scale.

Cite this page 2 min read

*Last updated on **May 2, 2026** by **Seid Muhie Yimam***

Running a Playbook Workshop

For larger groups or project kick-offs, a structured workshop helps align a team on guidelines and workflows before annotation begins.

When a Workshop Is Worth It

A dedicated session makes sense when:

- You are starting a new project with contributors who have not used the playbook before
- You need to align a large team on common terminology before annotation begins
- You are piloting the playbook in a new language community and want to collect localization feedback

For small teams or individual onboarding, a one-to-one walkthrough is usually sufficient — see Onboarding a Team.

Workshop Formats

Format	Best for	Duration
Orientation session	Introduce the playbook to a new team	2–3 hours
Task deep dive	Train contributors on one chapter (e.g., speech recording, text annotation)	Half day
Full onboarding	Bring all project roles together before a project launch	Full day
Feedback and localization	Review a chapter with community experts, collect revision input	2–3 hours

Preparing a Session

1. **Identify the relevant chapters** — limit to 1–2 per session; do not attempt the full playbook at once
2. **Prepare a concrete worked example** — use real samples from your project, not abstract examples
3. **Distribute materials in advance** — share the chapter link or a printout at least 3 days beforehand
4. **Assign roles** — designate a facilitator, a note-taker, and a timekeeper before the session begins
5. **Make materials available offline** — see Inclusive Access

A Simple Session Structure

```
00:00 – 00:15  Welcome and objectives
00:15 – 00:30  Playbook overview (structure, how to navigate)
00:30 – 01:00  Deep dive into the relevant chapter (facilitator-led)
01:00 – 01:30  Hands-on task in small groups (3–5 people)
01:30 – 01:50  Group debrief (what was clear, what was confusing)
01:50 – 02:00  Next steps and Q&A
```

For a full-day workshop, repeat the deep-dive and hands-on blocks for each additional chapter with breaks between.

Facilitation Tips

- Start with a concrete task, not a lecture — participants engage faster when they have something to do
- **Capture disagreements** — where participants interpret guidelines differently is where the playbook needs more clarity; document these moments and submit them as feedback
- Time-box discussions; have someone keep the session on schedule

After the Workshop

- Share a written summary within 48 hours: what was covered, what questions arose, what next steps are
- Log any sections that caused confusion — these are candidates for improvement
- Submit feedback via the GitHub repository or the built-in feedback form on the site
- Identify one or two participants who can serve as local playbook champions — people who can answer questions and onboard future contributors independently

Cite this page 2 min read

*Last updated on **May 2, 2026** by **Seid Muhie Yimam***

Inclusive Access

The playbook should be usable by people with different levels of technical background, literacy, language, and connectivity. This page guides coordinators on adapting it for diverse audiences.

Adapting for Non-Technical Audiences

Many contributors — annotators, voice recorders, translators — are not researchers. When introducing the playbook to them:

- Share only the sections relevant to their task; do not link to the full playbook
- Replace technical terms with plain-language equivalents, or walk through the Glossary first
- Demonstrate tasks live (screen share or in-person) before asking contributors to read independently
- Use concrete examples from the contributor's own language and domain

Adapting for Low-Literacy Contributors

- Convert key instructions into short numbered steps with screenshots or illustrations
- Offer oral walkthroughs as an alternative to written reading — a coordinator reads through the section and takes questions
- Prepare a printed quick-reference card (A5, one side) with the most important steps for the specific task
- Test materials with community members before finalizing — contributors with low literacy often surface genuine clarity problems that polished prose hides

Gender-Inclusive Facilitation

- Actively invite women and underrepresented participants to contribute examples and ask questions

- Use training examples that reflect diverse speakers, topics, and perspectives — avoid stereotyped scenarios
- Offer flexible session timing to accommodate caregiving responsibilities
- Track participation by gender and adjust facilitation if one group dominates

Multilingual Access

- Translate the sections relevant to your project before the training session — do not rely on machine translation for consent forms or contributor agreements
- Maintain a local glossary of key annotation terms in the community's working language
- Invite a bilingual co-facilitator for sessions where participants have limited English

Offline and Low-Connectivity Access

For contributors working without reliable internet:

- Use **Download PDF** in the navbar to get the full playbook as a single file — suitable for printing or sharing via WhatsApp or USB
- For workshops in low-connectivity venues, pre-download or print only the relevant chapter pages before traveling
- Ensure the annotation tool being demonstrated also supports offline or low-bandwidth use — see Tooling

When distributing printed copies, note the version number and date on the cover. The live site always reflects the latest version.

Collecting Feedback on Usability

After each training session, ask participants:

- Was there anything you did not understand after reading the relevant section?
- Were the examples relevant to your language and context?
- What would you change to make it easier to follow?

Submit responses to the playbook maintainers — the playbook improves through exactly this kind of community use.

Cite this page 2 min read

*Last updated on **May 2, 2026** by **Seid Muhie Yimam***

Cost and Resource Planning

Learn how to effectively plan the resources required for dataset creation, including budgeting, timelines, an...

Data Cleaning and Preprocessing

Learn how to prepare raw data for use in language AI systems by improving quality, consistency, and usabil...

Data Provenance and Traceability

Learn how to track the origin, history, and transformations of your data to ensure transparency, reproducibil...

Ethics, Bias, and Governance

Learn how to ensure responsible dataset creation by addressing bias, protecting privacy, and maintaining tr...

Cost and Resource Planning

Learn how to effectively plan the resources required for dataset creation, including budgeting, timelines, and scaling strategies.

Why Planning Matters

Dataset creation can be resource-intensive. Proper planning helps ensure efficient use of time, budget, and human effort while maintaining data quality.

Key Components of Planning

Budgeting Annotation Costs

- **Annotation cost estimation** – Calculate cost per sample or per task
- **Workforce planning** – Consider expert vs crowd annotators
- **Tooling costs** – Include platforms, storage, and infrastructure
- **Quality control costs** – Account for validation and review processes

Time Estimation

- **Task complexity** – More complex tasks require more time per annotation
- **Annotator speed** – Estimate based on pilot studies or benchmarks
- **Project phases** – Include setup, training, annotation, and validation
- **Buffer time** – Plan for delays and iterations

Scaling Strategy

- **Incremental scaling** – Start small and expand gradually
- **Automation support** – Use tools to speed up preprocessing and validation
- **Parallel workflows** – Distribute tasks across multiple annotators
- **Quality vs scale balance** – Maintain data quality while increasing size

*Last updated on **Apr 21, 2026** by **Tadesse Destaw***

Data Cleaning and Preprocessing

Learn how to prepare raw data for use in language AI systems by improving quality, consistency, and usability.

Why Data Cleaning Matters

Raw data often contains noise, inconsistencies, and errors. Proper cleaning and preprocessing ensure that datasets are reliable, accurate, and suitable for downstream tasks such as training and evaluation.

Key Steps in Data Cleaning

Deduplication, Normalization, and Filtering

- **Deduplication** – Remove duplicate entries to avoid bias and overrepresentation
- **Normalization** – Standardize text (e.g., casing, punctuation, encoding)
- **Filtering** – Remove irrelevant, low-quality, or out-of-scope data

Language Detection and Formatting

- **Language detection** – Identify and verify the language of each data instance
- **Formatting** – Ensure consistent structure (e.g., JSON, CSV, text fields)
- **Encoding consistency** – Maintain uniform character encoding (e.g., UTF-8)

Noise and Toxicity Handling

- **Noise removal** – Clean unwanted artifacts such as HTML tags, emojis (if not needed), or corrupted text
- **Toxicity handling** – Detect and manage harmful, offensive, or unsafe content depending on project goals

Missing and Corrupted Data Handling

- **Missing data** – Identify incomplete entries and decide whether to fill, ignore, or remove them
- **Corrupted data** – Detect broken or unreadable content and clean or discard it
- **Validation checks** – Ensure data integrity after preprocessing

 1 min read

*Last updated on **Apr 21, 2026** by **Tadesse Destaw***

Data Provenance and Traceability

Learn how to track the origin, history, and transformations of your data to ensure transparency, reproducibility, and accountability.

Why Provenance Matters

Understanding where data comes from and how it has been processed is essential for building trustworthy datasets. Provenance supports reproducibility, enables auditing, and helps identify potential issues in data quality and bias.

Key Components of Provenance

Source Tracking

- **URLs and references** – Record links or original sources of the data
- **Contributors** – Track who collected, created, or provided the data
- **Collection context** – Document when, where, and how the data was obtained

Data Lineage

- **Data evolution** – Track how data changes over time
- **Versioning** – Maintain different versions of datasets
- **Pipeline tracking** – Document each stage of data processing

Transformation Logs

- **Preprocessing steps** – Record cleaning, normalization, and filtering operations
- **Annotation processes** – Track labeling methods and guidelines used
- **Modifications** – Log any changes made to the data after collection
- **Audit trails** – Maintain records for reproducibility and verification

 1 min read

*Last updated on **Apr 21, 2026** by **Tadesse Destaw***

Ethics, Bias, and Governance

Learn how to ensure responsible dataset creation by addressing bias, protecting privacy, and maintaining transparency throughout the data lifecycle.

Why Ethics and Governance Matter

Datasets directly influence the behavior of language AI systems. Poor handling of bias, privacy, or sensitive content can lead to harmful outcomes. Ethical practices and governance frameworks help ensure fairness, accountability, and trust.

Key Components of Ethical Data Practices

Bias Identification and Mitigation

- **Bias detection** – Identify imbalances or skewed representations in data
- **Source bias** – Assess biases introduced by data sources
- **Annotation bias** – Monitor inconsistencies across annotators
- **Mitigation strategies** – Apply re-sampling, re-weighting, or guideline refinement

PII Detection and Removal

- **Personal data identification** – Detect names, addresses, contact details, and identifiers
- **Automated detection tools** – Use models or rules to flag sensitive information
- **Manual review** – Validate automated detection with human checks
- **Data removal or masking** – حذف or obfuscate personal identifiers

Anonymization Strategies

- **De-identification** – Remove or replace identifiable information
- **Pseudonymization** – Substitute identifiers with artificial labels
- **Aggregation** – Present data in grouped form to prevent re-identification
- **Risk assessment** – Evaluate re-identification risks after anonymization

Sensitive Attribute and Content Handling

- **Sensitive attributes** – Gender, ethnicity, religion, health, or political views
- **Content moderation** – Handle harmful, offensive, or explicit content carefully
- **Access control** – Restrict sensitive data to authorized users
- **Use-case alignment** – Decide inclusion based on task requirements

Fair Representation

- **Inclusive sampling** – Ensure diverse representation across groups
- **Balanced datasets** – Avoid over- or under-representation
- **Context awareness** – Consider cultural and linguistic diversity
- **Evaluation fairness** – Test models across different subgroups

Risk Documentation and Transparency

- **Risk identification** – Document potential harms and limitations
- **Datasheets and documentation** – Provide clear dataset descriptions
- **Transparency practices** – Share collection, processing, and annotation details
- **Governance policies** – Define rules for dataset usage and distribution

 2 min read

*Last updated on **Apr 21, 2026** by **Tadesse Destaw***

Annotation Task Design and Human Factors

Task complexity estimation

Inclusive and Bias-Aware Annotation

Learn how to design annotation processes that are inclusive, fair, and aware of potential biases introduced ...

Training and Guidelines

Learn how to prepare annotators through clear instructions, structured training, and iterative refinement of ...

Workflow and Adjudication

Multi-annotator setup

Annotation Task Design and Human Factors

Task complexity estimation

Annotator fatigue

UI/UX considerations

Cite this page 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Inclusive and Bias-Aware Annotation

Learn how to design annotation processes that are inclusive, fair, and aware of potential biases introduced by annotators and data.

Why Inclusion and Bias Awareness Matter

Annotation is a human-driven process, and annotator backgrounds can influence labeling decisions. Ensuring diversity and awareness helps reduce bias and improves the quality and fairness of datasets.

Key Components

Gender, Age, and Cultural Diversity

- Include annotators from diverse **gender**, **age groups**, and **cultural backgrounds**
- Ensure representation across different **dialects and communities**
- Avoid over-reliance on a single demographic group
- Consider cultural context when interpreting data

Recording Annotator Personas

- Document annotator characteristics where appropriate and ethical
- Capture information such as **language background**, **region**, or **expertise**
- Use anonymized metadata to analyze potential annotation biases
- Ensure privacy and consent when collecting annotator information

Bias Awareness Training

- Train annotators to recognize **personal and cultural biases**
- Provide examples of biased vs unbiased annotations

- Encourage consistent application of guidelines
- Reinforce neutrality and objectivity in labeling

Community Participation

- Engage local communities in the annotation process
- Incorporate native speaker knowledge and cultural insights
- Promote participatory and inclusive dataset creation
- Respect community norms and values throughout the process

 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Training and Guidelines

Learn how to prepare annotators through clear instructions, structured training, and iterative refinement of annotation tasks.

Annotation Guidelines with Examples

Clear and consistent guidelines are the foundation of reliable annotation.

- Define each label or category in simple and unambiguous terms
- Provide **positive examples** for each label
- Provide **negative examples** to clarify boundaries
- Include **edge cases** to handle ambiguity
- Specify how to treat uncertain or mixed cases
- Use consistent formatting and terminology throughout the guideline

Training and Calibration Rounds

Training ensures annotators understand and apply guidelines consistently.

- Conduct initial training sessions before annotation begins
- Use calibration tasks to align annotator understanding
- Compare annotations across multiple annotators for the same samples
- Provide structured feedback to resolve misunderstandings
- Repeat calibration until acceptable agreement is reached

Pilot Annotation and Iteration

Pilot annotation helps test and refine the annotation design before scaling.

- Start with a small subset of data
- Identify unclear instructions or confusing labels
- Measure annotation consistency and difficulty

- Collect annotator feedback on task clarity
- Iteratively refine guidelines, labels, and workflow

Minimum Viable Dataset per Task

A minimum viable dataset ensures the task design is valid before full-scale annotation.

- Create a small but representative dataset for each task
- Validate label schema coverage and clarity
- Test annotation workflow and tool usability
- Check feasibility of large-scale annotation
- Use results to decide whether to scale or redesign the task

Cite this page 2 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Workflow and Adjudication

Multi-annotator setup

Task assignment and redundancy

Disagreement resolution and expert adjudication

Cite this page 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Data Quality Management

Class imbalance handling

Data Quality Assurance

Inter-Annotator Agreement (Kappa, Alpha)

Data Quality Management

Class imbalance handling

Outlier and noise detection

Error analysis pipelines

Dataset versioning and updates

Cite this page 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Data Quality Assurance

Inter-Annotator Agreement (Kappa, Alpha)

Gold data and control checks

Auditing and spot-checking

Feedback loops

 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Image Data

- Image Classification and Recognition – (object classification, scene recognition)

Multimodal Data

- Vision-Language Tasks – (image-text retrieval, captioning, VQA)

Speech Data

- ASR (Automatic Speech Recognition) – (transcription, multilingual ASR, code-switching)

Text Data

- Text Classification – (sentiment, emotion, hate speech, topic, intent detection)

Image Data

- **Image Classification and Recognition** – (object classification, scene recognition)
- **Object Detection and Segmentation** – (bounding boxes, instance/semantic segmentation)
- **Image Captioning and Generation** – (captioning, image-to-text generation)
- **Vision-Language Tasks** – (Visual Question Answering (VQA), referring expressions)
- **Image-to-Image Tasks** – (style transfer, super-resolution, restoration)
- **Document Understanding** – (OCR, layout analysis, form understanding)

 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Multimodal Data

- **Vision-Language Tasks** – (image-text retrieval, captioning, VQA)
- **Audio-Text Tasks** – (speech translation, audio captioning)
- **Cross-Modal Retrieval and Alignment** – (text-to-image, image-to-text search)
- **Multimodal Generation** – (text-to-image, text-to-video, image-conditioned text generation)
- **Reasoning and Instruction Tasks** – (multimodal QA, instruction following)

 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Speech Data

- **ASR (Automatic Speech Recognition)** – (transcription, multilingual ASR, code-switching)
- **TTS (Text-to-Speech)** – (single-speaker, multi-speaker, expressive TTS)
- **Speech-to-Speech Translation (STS)** – (direct speech translation across languages)
- **Audio Understanding** – (audio classification, sound event detection)
- **Speech emotion recognition**
- **Speaker diarization**

Cite this page 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Text Data

- **Text Classification** – (sentiment, emotion, hate speech, topic, intent detection)
- **Sequence Labeling** – (NER, POS tagging, chunking, slot filling, keyphrase extraction)
- **Sequence-to-Sequence** – (machine translation, summarization, paraphrasing, simplification)
- **Question Answering and Reasoning** – (extractive QA, generative QA, reading comprehension)
- **Retrieval and Ranking** – (document retrieval, semantic search, reranking)
- **Dialogue and Generation** – (chatbots, instruction following, story generation)
- **Structured Prediction and Parsing** – (dependency parsing, constituency parsing, semantic parsing)

 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Templates & Artifacts

This page lists all governance toolkit templates available in the playbook. Each template is a standalone doc...

Consent Form Template

When to Use This Template

Contributor Agreement Template

When to Use This Template

Data Ownership Documentation Template

When to Use This Template

Licensing and Compliance

Common License Types for NLP Datasets

Sustainability Plan Template

When to Use This Template

📄 Documentation and Reporting

- Datasheets for datasets

📄 Tooling and Infrastructure

- Annotation tools (usability, scalability, cost)

📄 Data Storage and Release Infrastructure

- Repository hosting (Hugging Face, GitHub, institutional)

Templates & Artifacts

This page lists all governance toolkit templates available in the playbook. Each template is a standalone document you can download, adapt, and use directly in your project.

Template	Who uses it	When
Consent Form	Project leads, coordinators	Before data collection begins
Contributor Agreement	Project leads	When onboarding annotators or chapter authors
Data Ownership Documentation	Project leads, legal reviewers	At project setup and dataset release
Sustainability Plan	Project leads, funders	At project design and end-of-grant stage
Licensing Agreement	Project leads, legal reviewers	Before dataset publication

Other Artifacts

- Annotation guideline template → see Training and Guidelines
- Datasheet template → see Documentation and Reporting
- Data collection plan template → see Cost and Resource Planning

Cite this page 1 min read

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Consent Form Template

When to Use This Template

What This Form Covers

Part 1 — Project Information

Part 2 — What You Are Agreeing To

Part 3 — Data Use and Storage

Part 4 — Your Rights

Part 5 — Signature

Adapting for Low-Literacy Contexts

Adapting for Oral or Audio Consent

1 min read

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Contributor Agreement Template

When to Use This Template

What This Agreement Covers

Part 1 — Contributor and Project Details

Part 2 — Scope of Contribution

Part 3 — Intellectual Property and Licensing

Part 4 — Compensation and Attribution

Part 5 — Confidentiality and Data Handling

Part 6 — Termination

Part 7 — Signature

1 min read

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Data Ownership Documentation Template

When to Use This Template

What This Document Covers

Part 1 — Dataset Identification

Part 2 — Ownership and Rights Holders

Part 3 — Source Data and Third-Party Rights

Part 4 — Contributor Contributions

Part 5 — Licensing and Permitted Uses

Part 6 — Restrictions and Obligations

Part 7 — Contact and Dispute Resolution

1 min read

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Licensing and Compliance

Common License Types for NLP Datasets

Choosing the Right License

Restrictions and Attribution Requirements

Legal and Ethical Compliance

Licensing Agreement Template

Part 1 — Dataset and Licensor Details

Part 2 — Grant of Rights

Part 3 — Restrictions

Part 4 — Attribution Requirements

Part 5 — Warranty and Liability

Part 6 — Termination

Part 7 — Signature

 1 min read

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Sustainability Plan Template

When to Use This Template

What This Plan Covers

Part 1 — Project and Dataset Overview

Part 2 — Stewardship and Governance After the Grant

Part 3 — Hosting and Access

Part 4 — Community Maintenance and Update Process

Part 5 — Translation and Localization Pipeline

Part 6 — Funding and Resource Strategy Beyond the Grant

Part 7 — Risk and Contingency

Part 8 — Milestones and Review Schedule

 1 min read

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Documentation and Reporting

- Datasheets for datasets
- Standard reporting and reproducibility
- Failure cases and limitations
- Transparency in dataset creation

Cite this page 1 min read

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Tooling and Infrastructure

- Annotation tools (usability, scalability, cost)
- Data pipelines and automation
- Deployment (cloud vs local)
- Security and access control

Cite this page 1 min read

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Data Storage and Release Infrastructure

- Repository hosting (Hugging Face, GitHub, institutional)
- File formats and metadata standards
- Versioning and changelogs

Cite this page 1 min read

*Last updated on **May 3, 2026** by **Seid Muhie Yimam***

Synthetic Data Creation

- Data augmentation - (paraphrasing, back-translation)

LLM-Assisted Annotation

- LLM-assisted annotation - (human-in-the-loop)

Synthetic Data Creation

- **Data augmentation** - (paraphrasing, back-translation)
- **Fully synthetic generation** - (LLMs, simulation)
- **Scenario-based data generation**
- **Validation against real-world distributions**
- **Evaluation of synthetic data quality**

Cite this page 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

LLM-Assisted Annotation

- **LLM-assisted annotation** - (human-in-the-loop)
- **LLM and a human annotator agreement**
- **Prompt design and output validation**
- **Bias, hallucination, and consistency control**
- **When NOT to use LLMs**
- **Cost vs quality comparison** (LLM vs human)

Cite this page 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Data Integrity and Contamination Control

- Preventing train-test leakage

Evaluation, Benchmarking, and Data Integrity

- Evaluation Metrics by task

Data Integrity and Contamination Control

- Preventing train-test leakage
- Overlap with existing benchmarks
- LLM contamination (training data exposure)

Cite this page 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Evaluation, Benchmarking, and Data Integrity

- Evaluation Metrics by task
- Train/dev/test splits
- Cross-lingual and domain generalization
- Bias and robustness evaluation
- Bias evaluation metrics

Model Building and Starter Kits

- Baseline models for each modality/task
- Training and evaluation scripts
- Reproducibility guidelines
- Benchmark leaderboards
- Benchmark positioning

 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Maintenance and Post-Release Strategy

Dataset updates and versioning policy

Release Checklist

Data cleaned and validated

Maintenance and Post-Release Strategy

Dataset updates and versioning policy

Deprecation strategy

Community feedback loops

Issue tracking

Cite this page 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Release Checklist

Data cleaned and validated

Annotation quality verified

Documentation completed

Licensing defined

Ethical review conducted

Baselines and splits provided

Public access ensured

Cite this page 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Collaboration and Shared Tasks

Shared tasks and benchmarks

Community Ecosystems

Community initiatives (Masakhane, EthioNLP, HausaNLP)

Collaboration and Shared Tasks

Shared tasks and benchmarks

Workshops and open challenges

Cite this page 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Community Ecosystems

Community initiatives (Masakhane, EthioNLP, HausaNLP)

Academic and industry collaboration

Contribution and contributor guidelines

Cite this page 1 min read

*Last updated on **Apr 22, 2026** by **Tadesse Destaw***

Defining text classification tasks:

Text classification is one of the most fundamental tasks in Natural Language Processing (NLP). The goal is ...

Data sources

Common Data Sources: Selecting data sources that are relevant, ethical, and representative is essential for ...

Data Collection and Selection Approaches

Text data can be collected through APIs, web scraping (with permission), manual collection, or surveys, wh...

Data Processing and Sampling

Once the data source has been identified, several preprocessing and sampling steps are required to ensure...

Annotation Tools

Text classification data can be annotated using a range of tools, from managed crowdsourcing platforms to...

Annotator Recruitment/Selection

Effective annotation relies more on the quality and consistency of annotators than on their number. Annotat...

Annotation Quality Control

Annotation quality can be controlled before and during the annotation process using various mechanisms. ...

Annotation Agreement

Annotation agreement measures the extent to which multiple annotators assign the same labels to the sam...

Sentiment Analysis

The Sentiment Analysis Annotation Guidelines Template is a structured framework designed to standardize ...

Emotion Analysis

What is Emotion Analysis?

Hate Speech Analysis

What is Hate Speech Analysis?

Data Quality Control

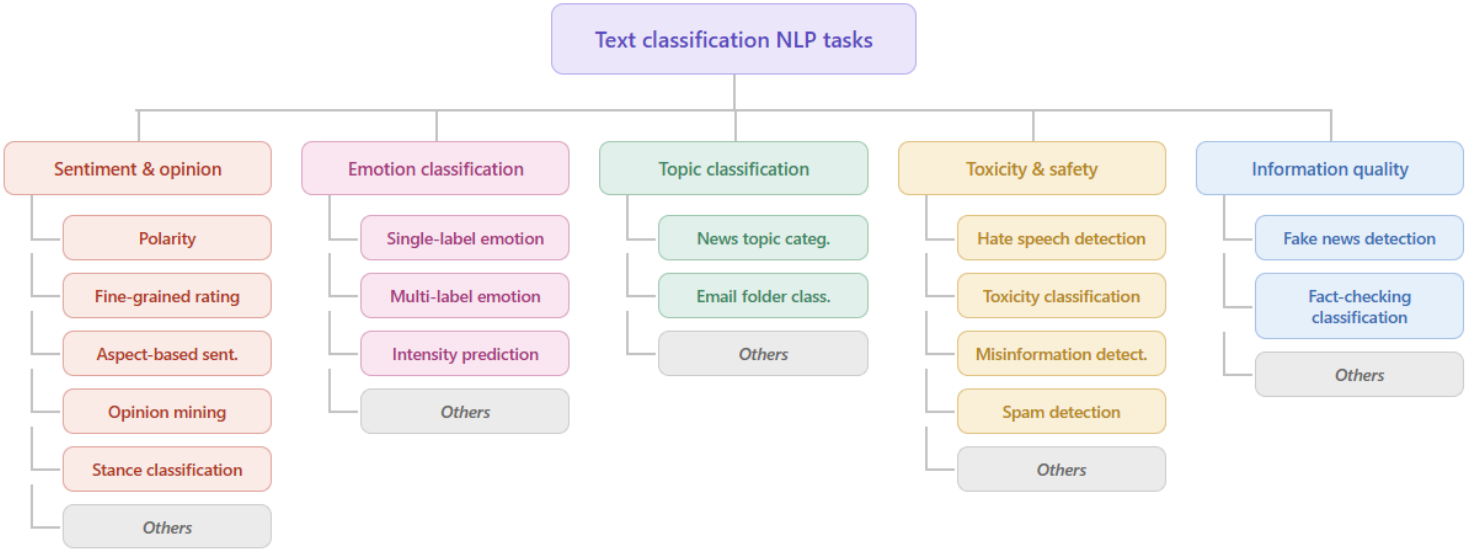
Data quality control ensures that your sentiment labels are accurate, consistent, and reliable.

Annotator Safety and Mental Health

Protecting annotator well-being is essential, particularly when working with harmful, offensive, or emotiona...

Defining text classification tasks:

Text classification is one of the most fundamental tasks in Natural Language Processing (NLP). The goal is to automatically assign one or more predefined labels (categories) to a piece of text. Among numerous text classification NLP tasks, sentiment analysis, hate speech classification, emotion classification, and topic classification are the most common. In this guidebook, we will discuss details about these common NLP tasks. While there is no single agreed-upon definition of the following NLP tasks, we use the most widely agreed-upon definitions.

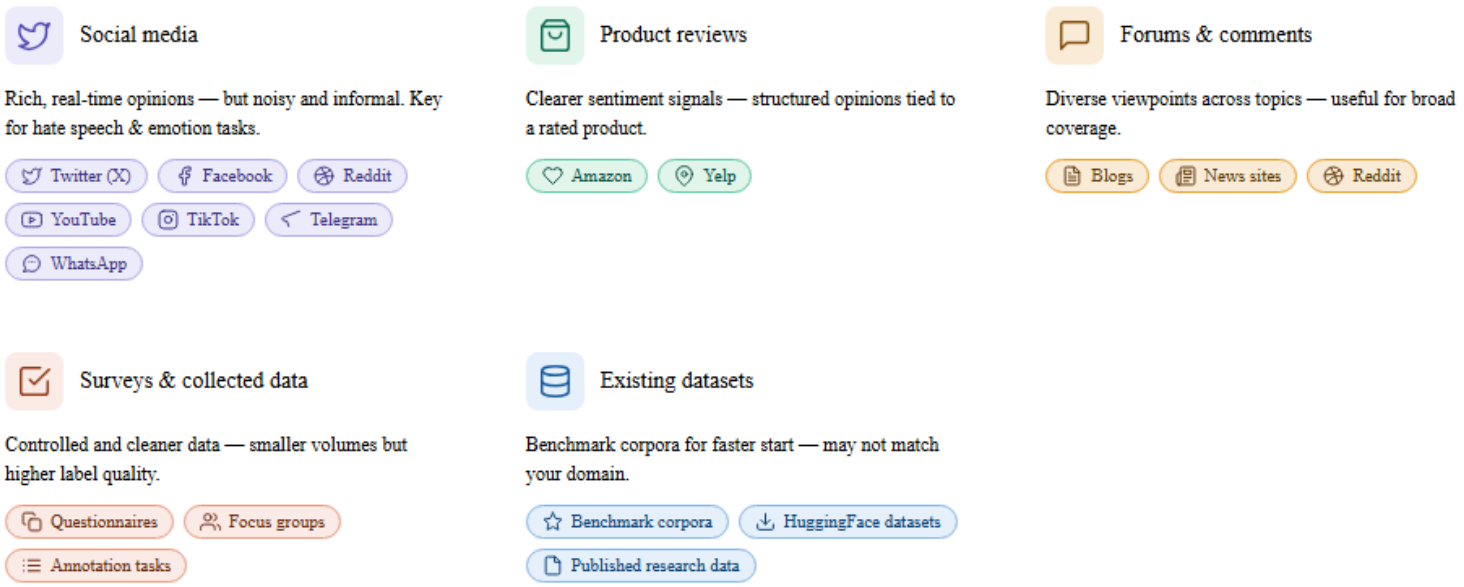


Cite this page

*Last updated on **Jun 9, 2026** by **Tadesse Destaw Belay***

Data sources

Common Data Sources: Selecting data sources that are relevant, ethical, and representative is essential for building high-quality text classification datasets such as sentiment, emotion, and hate speech datasets. Common sources include social media platforms such as Twitter (X), Facebook, Reddit, YouTube, TikTok, Telegram, and WhatsApp, which provide rich and real-time user opinions but often contain noisy and informal language. Product review platforms such as Amazon typically offer clearer sentiment signals, while forums, blogs, and news comment sections provide diverse viewpoints and discussions. Researchers may also collect data through surveys or controlled studies, which generally produce cleaner but smaller datasets. Additionally, existing benchmark datasets can accelerate research and enable comparison with prior work, although they may not always align with the target domain, language, or cultural context.



Cite this page

*Last updated on **Jun 9, 2026** by **Tadesse Destaw Belay***

Data Collection and Selection Approaches

Text data can be collected through APIs, web scraping (with permission), manual collection, or surveys, while preserving useful metadata such as source, time, language, and identifiers for future analysis. Data sources should be relevant to the target domain, language, and cultural context, with careful attention to dataset quality, class balance, and representativeness. Throughout the process, researchers must also address ethical and legal requirements, including privacy, consent, and compliance with platform policies. Data samples can be collected using one of the approaches below.

Keyword/Dictionary-based Selection: Select documents or sentences that contain one or more predefined keywords, phrases, or lexicon entries. For example, in hate speech detection, a list of commonly used hate-related or offensive terms in the target language can be compiled and used to identify potentially relevant texts. This method helps enrich the dataset with task-relevant examples while reducing the amount of irrelevant data. A widely used resource for English emotion-related keywords is the NRC Emotion Lexicon.

PROS	CONS
✓ Highly targeted samples	✗ Misses paraphrases & synonyms
✓ No model needed to select	✗ Selection bias toward keywords
✓ Interpretable & auditable	✗ Requires domain knowledge

Location-based Selection: Is a data collection approach where texts are gathered based on the geographic location associated with users or posts. For example, when collecting social media data from X (formerly Twitter), researchers can filter posts originating from specific locations such as Ethiopia, Nigeria, or Kenya. This method is useful for studying regional language variation, local opinions, cultural expressions, or location-specific events, as it helps ensure that the collected data represents the target geographic area.

PROS

- ✓ Enables dialect / regional studies
- ✓ Precise targeting for geo tasks
- ✓ Scales easily with bounding boxes

CONS

- ✗ Only ~1–3% of tweets are geotagged
- ✗ Location ≠ author's language region
- ✗ Platform-dependent metadata

Distant supervision: Is a method for automatically creating labeled training data by using existing knowledge sources instead of manual annotation. For example, for emotion classification, social media posts containing hashtags such as **#happy**, **#joy**, or **#sad** can be automatically labeled with the corresponding emotions. This approach enables the creation of large training datasets quickly and cheaply, although some automatically assigned labels may be incorrect or noisy.

PROS

- ✓ Millions of examples, no manual labels
- ✓ Enables training without annotation budget
- ✓ Can bootstrap active-learning cycles

CONS

- ✗ Label noise can hurt model quality
- ✗ Assumes signal aligns with task
- ✗ Needs noise-aware training strategy

Random Sampling: Select items uniformly at random from the corpus with no targeting criteria. This is the baseline for any annotation project, ensuring an unbiased estimate of the true corpus-level label distribution.

PROS

- ✓ Zero selection bias — most representative
- ✓ Simple to implement and audit
- ✓ Best for estimating true distributions

CONS

- ✗ Rare classes underrepresented
- ✗ Wastes budget on uninformative samples
- ✗ Inefficient for imbalanced tasks

Active Learning Method: Iteratively train a model on a small seed set, then use the model's uncertainty to select the most informative unlabelled examples for human annotation next. Maximizes annotation return on investment by labeling only where the model is confused.

PROS

- ✓ Up to 70% fewer labels for same accuracy
- ✓ Naturally surfaces hard edge cases
- ✓ Improves model and data simultaneously

CONS

- ✗ Requires annotation loop infrastructure
- ✗ Bias toward uncertain decision boundary
- ✗ Cold-start: needs an initial labelled seed

Stratified Sampling: Divide the corpus into strata — subgroups by class, source, time period, or demographic — and sample proportionally or equally from each. Ensures minority classes and subgroups are always represented in the annotation set.

PROS

- ✓ Guarantees coverage of rare classes
- ✓ Controls for known confounders
- ✓ Produces balanced training sets

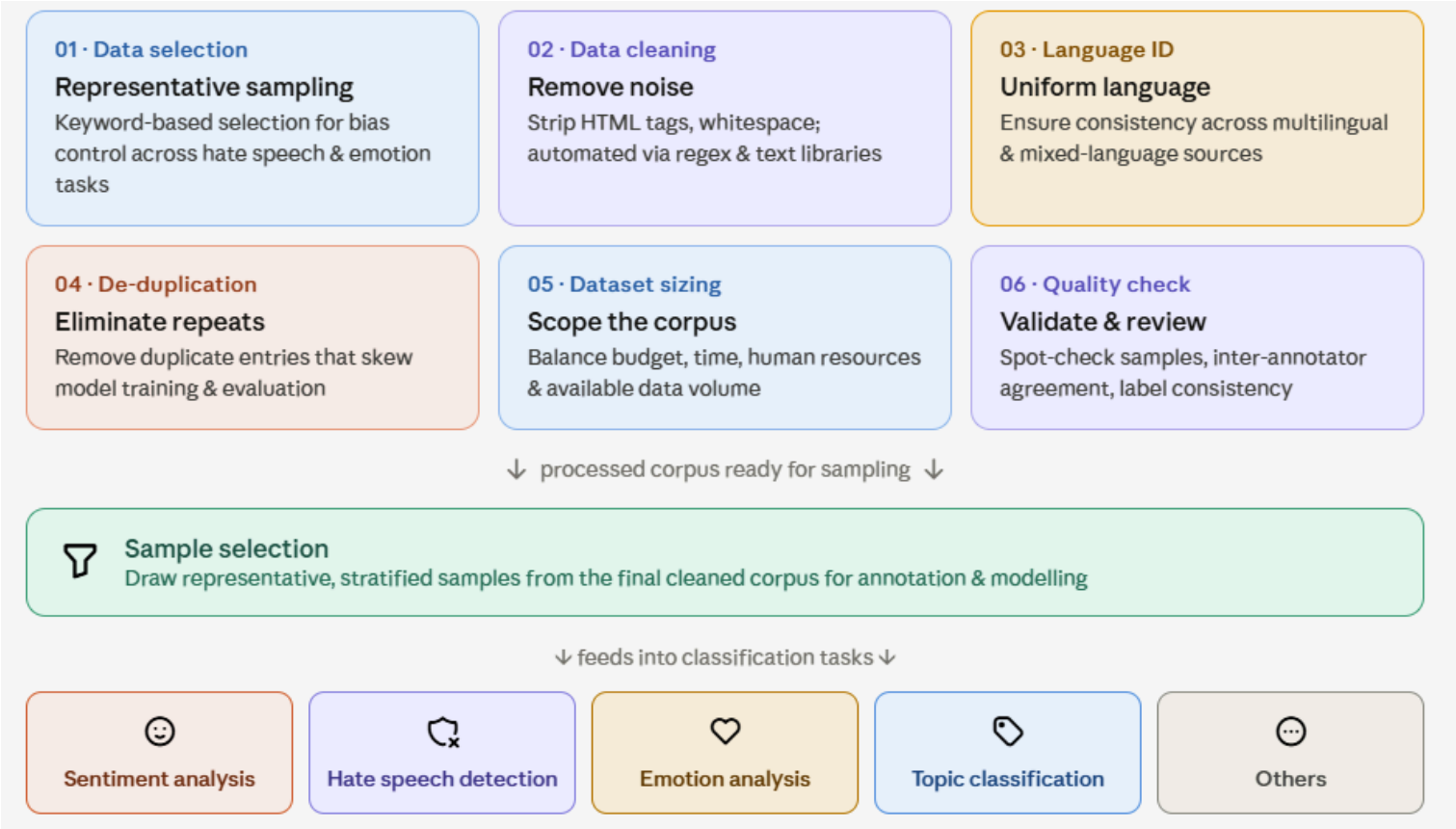
CONS

- ✗ Requires pre-existing strata metadata
- ✗ Equal sampling distorts true distribution
- ✗ Strata boundaries can be arbitrary

*Last updated on **Jun 9, 2026** by **Tadesse Destaw Belay***

Data Processing and Sampling

Once the data source has been identified, several preprocessing and sampling steps are required to ensure the quality and representativeness of the dataset. First, texts should be carefully selected to reflect the diversity of the target population and minimize potential biases. For tasks such as hate speech and emotion analysis, keyword-based filtering can be useful for identifying relevant content. Data cleaning involves removing irrelevant elements such as HTML tags, URLs, special characters, and excessive whitespace using text-processing tools. Applying language identification and de-duplication helps eliminate non-target language and repeated content. The overall dataset size should be determined based on factors such as research objectives, available resources, annotation budget, human capacity, and project timelines.



Cite this page

*Last updated on **Jun 9, 2026** by **Tadesse Destaw Belay***

Annotation Tools

Text classification data can be annotated using a range of tools, from managed crowdsourcing platforms to self-hosted open-source systems. The choice of tool should depend on the task design, dataset size, number of annotators, required turnaround time, and the availability of qualified annotators.

There is no single best tool. When selecting an annotation tool, the right choice depends on several factors:

- Data sensitivity and privacy requirements.
- Size of the dataset and expected annotation volume.
- Number of annotators and required collaboration features.
- Need for quality assurance, audit trails, and reviewer workflows.
- Support for the annotation schema and task type.
- Availability of self-hosting, access control, and export options.
- Cost, ease of setup, and long-term maintainability.

Based on how annotators are sourced and where the data is hosted, annotation tools can be broadly grouped into two categories: crowdsourcing platforms and in-house (self-hosted) tools. The most common options for text classification are described below.

5.1. Crowdsourcing Platforms

Crowdsourcing platforms give you access to a large, on-demand pool of remote annotators who are recruited and paid through the platform. Annotators are typically selected based on profile attributes — language, location, demographics, prior approval rating, or qualification tests — rather than being known to you personally. This makes crowdsourcing well suited to large volumes of data and widely spoken languages, where a broad annotator pool is readily available.

Common platforms include:

- Toloka AI — https://toloka.ai
- Amazon Mechanical Turk (MTurk) — https://www.mturk.com
- Prolific — https://www.prolific.com

- Appen — https://www.appen.com
- Label Studio Enterprise — https://labelstud.io

Many of these platforms support multiple data modalities (text, image, audio, video) and increasingly offer AI-assisted features such as pre-labeling, model-in-the-loop suggestions, and automated quality checks.

Trade-offs. Crowdsourcing scales easily and reduces recruitment overhead, but it offers less direct control over annotators and may make it difficult to source native speakers of low-resource languages or specific dialects. It also requires careful task design, qualification filters, and it raises data-privacy considerations because the data is sent to an external platform.

5.2. In-House (Self-Hosted) Tools

In-house tools are typically open-source applications that can be customized, deployed, and hosted on your own machine or server. You create accounts for a hand-picked set of annotators — often colleagues, domain experts, or recruited native speakers — giving you full control over who labels the data and where the data lives. This category is preferred for sensitive data, specialized domains, and low-resource languages, where annotator expertise matters more than raw scale.

Common self-hosted tools include:

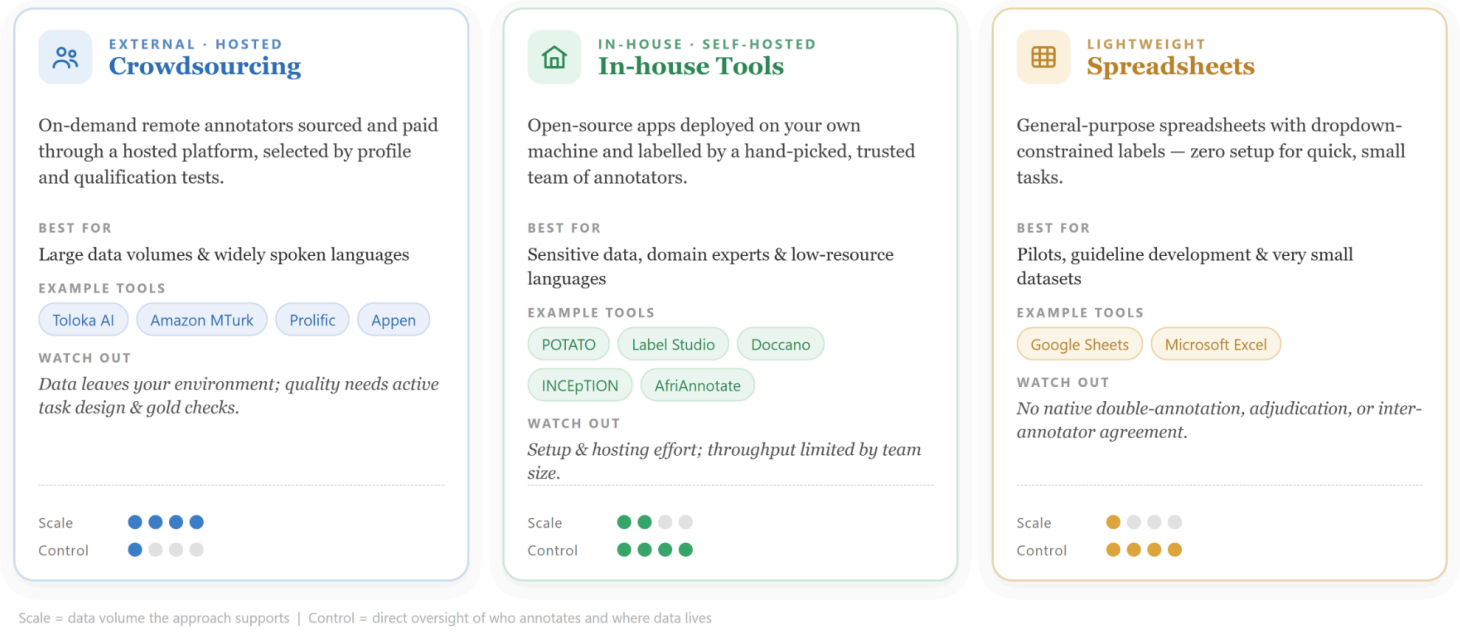
- POTATO — Portable Text Annotation Tool — https://github.com/davidjurgens/potato
- Label Studio (open-source edition) — https://labelstud.io
- Doccano — https://github.com/doccano/doccano
- INCEpTION (the successor to WebAnno) — https://inception-project.github.io
- brat — https://brat.nlplab.org

Trade-offs. Self-hosted tools keep data fully under your control and can be tailored to bespoke label schemes and guidelines, but they require setup, hosting, and maintenance effort, and the annotation throughput is limited by the size of your recruited team.

5.3. Lightweight Tools for Small Datasets

For small annotation efforts, a dedicated platform may be unnecessary. Spreadsheets — Google Sheets or Microsoft Excel — are a practical, zero-setup option: one column holds the text, and one or more columns capture the label(s), with data validation or dropdown lists used to constrain inputs to the allowed label set. Spreadsheets are easy to share and require no technical onboarding, which makes them convenient for pilot studies, guideline development, and very small in-house tasks.

However, these tools lack the quality-control and management features; they are not recommended beyond small or exploratory datasets.

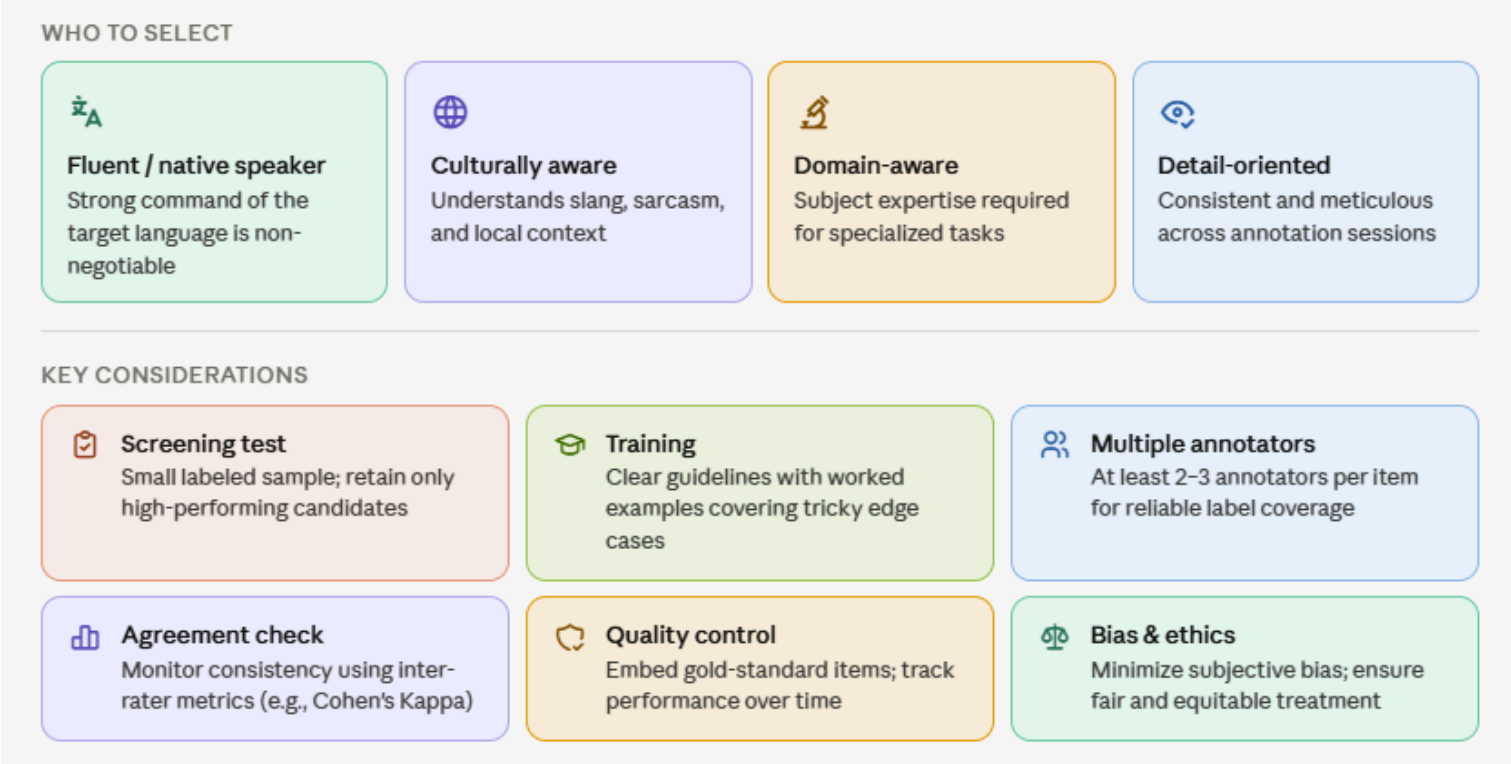


Cite this page

*Last updated on **Jun 9, 2026** by **Tadesse Destaw Belay***

Annotator Recruitment/Selection

Effective annotation relies more on the quality and consistency of annotators than on their number. Annotators should be fluent in the target language, familiar with the relevant cultural context, and, when necessary, possess domain-specific knowledge. They should also be detail-oriented and able to apply annotation guidelines consistently to ensure reliable, high-quality labels.



Cite this page

*Last updated on **Jun 9, 2026** by **Tadesse Destaw Belay***

Annotation Quality Control

Annotation quality can be controlled before and during the annotation process using various mechanisms. The following are some of the annotation quality control methods.

Pilot Annotation

Before starting large-scale annotation, dataset creators should conduct a pilot study to evaluate both the annotation guidelines and the annotators. The pilot phase helps identify unclear instructions, difficult cases, and annotators whose labeling patterns differ substantially from the rest of the group. Annotators who consistently provide random, low-quality, or highly inconsistent annotations should be identified and excluded before the main annotation process begins.

Control (Gold-Standard) Questions

A common quality-control mechanism is to include control questions, also known as gold-standard items, whose correct labels are already known. These items are randomly inserted into the annotation workflow without informing the annotators. Annotators who repeatedly fail to label these control items correctly may be removed from the project, and their previously annotated data should be reviewed and, if necessary, excluded from the final dataset.

Determining the Number of Annotators

For most NLP annotation tasks, using at least three annotators per instance is a common practice for ensuring annotation quality. An odd number of annotators (e.g., 3, 5, or 7) enables majority voting to determine the final label. In general, increasing the number of annotators per instance improves the reliability and robustness of the dataset by reducing the impact of individual biases. However, annotation cost and annotator availability often limit the number of annotators that can be employed.

When human resources are limited, annotation can be performed by two annotators. In such cases, dataset creators may either retain only the instances on which both annotators agree or introduce an adjudication process, where disagreements are resolved through discussion or by an expert annotator who makes the final decision.

 2 min read

*Last updated on **Jun 9, 2026** by **Tadesse Destaw Belay***

Annotation Agreement

Annotation agreement measures the extent to which multiple annotators assign the same labels to the same data instances. In text classification tasks, agreement is one of the most important indicators of dataset quality because it reflects the clarity of the annotation guidelines, the complexity of the task, and the consistency of the annotators. High agreement suggests that the labels are reliable and reproducible, while low agreement may indicate ambiguous definitions, insufficient annotator training, or inherently subjective phenomena.

Why Annotation Agreement Matters

Annotation agreement serves various purposes:

- Evaluates the reliability of the annotated dataset.
- Identifies ambiguities in the annotation guidelines.
- Detects inconsistencies among annotators.
- Provides evidence of dataset quality for publications and benchmark releases.
- Helps determine whether a task is objectively measurable or highly subjective.

Agreement should be calculated and reported for every dataset that involves human annotation if the data is annotated by two and more annotators.

Percentage Agreement

The simplest measure of agreement is percentage agreement, which calculates the proportion of instances for which annotators assigned the same label.

```
Agreement %= (Number of Agreed instances / Total Number of Instances) * 100

*For example, if two annotators label 1,000 texts and agree on 850 of them:*

Agreement % = (850 / 1000) * 100 = 85%
```

Although easy to understand, percentage agreement does not account for agreement occurring by chance and should not be the only metric reported.

Agreement Between Two Annotators

When exactly two annotators label each instance, Cohen's Kappa is the most commonly used agreement metric. Cohen's Kappa adjusts for the amount of agreement that could occur purely by chance.

```
Kappa = (Observed Agreement - Expected Agreement) / (1 - Expected Agreement)

Or

Kappa = (Po - Pe) / (1 - Pe)

Where:

Observed agreement (Po) is the proportion of instances where the annotators
actually agreed.

Po=Total number of items/Number of agreements

Expected Agreement (Pe) represents the level of agreement that would be
expected to occur purely by chance, given the distribution of labels
assigned by each annotator. It is calculated by determining the probability
that both annotators independently select the same category and then summing
these probabilities across all categories.
```

Cohen's Kappa is widely used in sentiment analysis, hate speech detection, topic classification, emotion classification, and many other NLP tasks involving two annotators.

Python Example

```
from sklearn.metrics import cohen*kappa*score

annotator1 = [0, 1, 1, 0, 2]

annotator2 = [0, 1, 0, 0, 2]
```

```
kappa = cohen*kappa*score(annotator1, annotator2)

print(kappa)
```

Agreement Among Three or More Annotators

Many NLP datasets use three or more annotators per instance to improve reliability and reduce the influence of individual biases.

When more than two annotators are involved, commonly used agreement measures include:

Fleiss' Kappa

Fleiss' Kappa extends Cohen's Kappa to multiple annotators and is one of the most widely reported agreement measures in NLP datasets.

It is appropriate when:d

- Three or more annotators label each instance.
- Every instance receives the same number of annotations.

```
Fleiss kappa (k) = P-Pe)/(1-Pe)

Where

p is the mean of the agreement probability over all raters and

Pe is the mean agreement probability over all raters if they were randomly
assigned.
```

Krippendorff's Alpha

Krippendorff's Alpha is a more flexible agreement measure that:

- Supports any number of annotators.
- Handles missing annotations.
- Works with nominal, ordinal, interval, and ratio labels.
- Is increasingly recommended for modern annotation studies.

For complex annotation projects, Krippendorff's Alpha is often considered the most robust agreement metric.

Deciding the Final Labels

When multiple annotators label the same instance, the final label is usually determined through majority voting.

For example, in three annotators, at least two annotators must agree on a label for it to become the final label. Similarly, with five annotators, at least three annotators must agree on a label for it to become the final label. Using an odd number of annotators (3, 5, or 7) avoids ties and simplifies majority voting.

Interpreting Agreement Scores

Although interpretation varies slightly across fields, the following ranges are commonly used for Kappa-based agreement measures:

Kappa Score Interpretation

```
< 0.00 Poor Agreement

0.00 - 0.20 Slight Agreement

0.21 - 0.40 Fair Agreement

0.41 - 0.60 Moderate Agreement

0.61 - 0.80 Substantial Agreement

0.81 - 1.00 Almost Perfect / Excellent Agreement

*As a general guideline:*
```

```
< 0.40: dataset quality should be carefully reviewed.

0.40-0.60: acceptable for difficult or subjective tasks.

0.60-0.80: considered good agreement.

Above 0.80: considered very strong agreement.
```

For highly subjective tasks such as emotion classification, sarcasm detection, or offensiveness annotation, lower agreement scores may still be acceptable due to genuine differences in human interpretation.

What to Report

When publishing a dataset, researchers should report:

1. Number of annotators.
2. Annotation procedure/giudeline.
3. Final label aggregation method (e.g., majority voting).
4. Cohen's Kappa (for two annotators) or Fleiss' Kappa/Krippendorff's Alpha (for three or more annotators) agreement score.
5. Any adjudication process used to resolve disagreements.
6. Annotator-level dataset for further annotator subjectivity and disagreement research.

Transparent reporting of annotation agreement improves the credibility, reproducibility, and scientific value of the dataset.

 4 min read

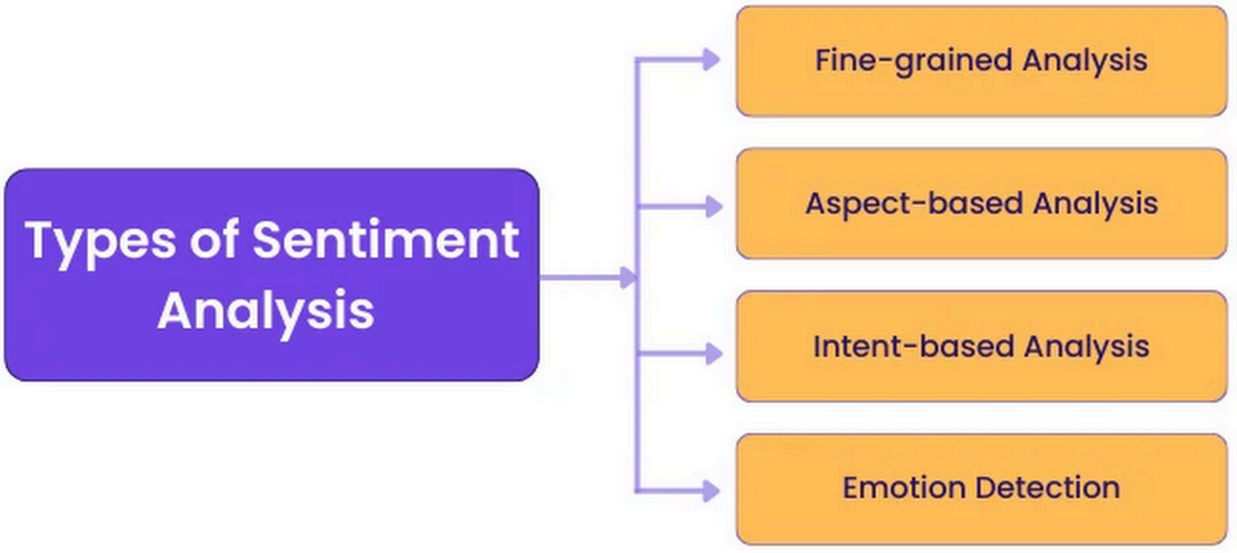
*Last updated on **Jun 9, 2026** by **Tadesse Destaw Belay***

Sentiment Analysis

The Sentiment Analysis Annotation Guidelines Template is a structured framework designed to standardize the process of annotating textual data for sentiment analysis. This template is particularly crucial in industries like customer service, marketing, and social media analytics, where understanding the sentiment behind user-generated content is essential. By providing clear instructions and examples, the template ensures that annotators can consistently label data as positive, negative, or neutral. For instance, in a customer feedback analysis project, this template helps teams align on how to interpret ambiguous phrases, ensuring the dataset's reliability and accuracy. The importance of such a template cannot be overstated, as it directly impacts the quality of machine learning models trained on the annotated data.

8.1. Types of sentiment analysis

Sentiment analysis can be annotated into one or more of the following annotation types.



**Fine-grained (Scaling) analysis: **Graded sentiment analysis assigns scores on a scale, providing a more nuanced view of sentiment intensity. This approach helps gauge the strength of emotions expressed in the text. For example, a review might be labeled as “very positive,” “slightly positive,” “neutral,” “slightly negative,” or “very negative.”

**Aspect-based analysis: **This focuses on identifying sentiment towards specific aspects or features of a product, service, or topic. For instance, in a hotel review, aspect-based analysis might determine positive sentiment towards the location but negative sentiment towards the cleanliness. For a smartphone review, it separately analyzes battery, screen, camera and performance to understand customer sentiment for each aspect.

Intent-based analysis

This analysis type can detect the motive behind a text, whether the author is trying to express an opinion, make a recommendation, ask a question, or express a need. Understanding intention is important in customer service, market research, and targeted advertising. For example, a customer tweets, 'I wish Company X's product had a longer battery life.' This indicates dissatisfaction and a desire for improvement (intent to recommend a feature change). This helps Company X handle the negativity and use this feedback to improve their products.

8.2. Sentiment analysis labels definition

While the types of sentiment targets can differ, the annotation labels can be two (positive and negative), three (positive, negative, and neutral), or all of the labels below.

- **Positive**: a message that conveys a clearly favorable attitude, such as satisfaction, approval, praise, or happiness toward a subject, product, or experience.
- **Negative**: a message that conveys a clearly unfavorable attitude, such as dissatisfaction, disapproval, criticism, or frustration toward a subject, product, or experience.
- **Mixed**: a message that includes both positive and negative sentiments, either about different aspects of the same subject or within the same statement.
- **Neutral**: a message that expresses no clear positive or negative sentiment, typically presenting factual, descriptive, or objective information without emotional judgment.

Examples of sentiment analysis are as follows:

Positive

- I love this car.
- This view is amazing.

- I feel great this morning.
- I am so excited about the concert.
- He is my best friend.

Negative

- I do not like this car.
- This view is horrible.
- I feel tired this morning.
- I am not looking forward to the concert.
- He is my enemy.

Mixed

- She is beautiful, but notorious.

Neutral

- There is a book on the desk.
- The sun lays on the sky.

Cite this page

*Last updated on **Jun 9, 2026** by **Tadesse Destaw Belay***

Emotion Analysis

What is Emotion Analysis?

Emotion detection or emotion classification is a task that aims to identify and classify the emotions expressed in a piece of text. Unlike sentiment analysis, which usually categorizes text as positive, negative, or neutral, emotion analysis seeks to recognize more specific emotional states such as anger, joy, sadness, fear, surprise, disgust, and neutral.

Emotion annotation goes a step further than sentiment analysis by analyzing specific emotions and classifying the text into categories such as joy, anger, sadness, fear, or surprise. By recognizing these, companies can better understand customer responses, which helps them respond to specific issues accordingly. For example, if you can identify any frustration in a customer complaint, you can address the issue immediately and prevent escalation. Emotion can be annotated into one of the following approaches.

9.1. Single Label Emotion Analysis

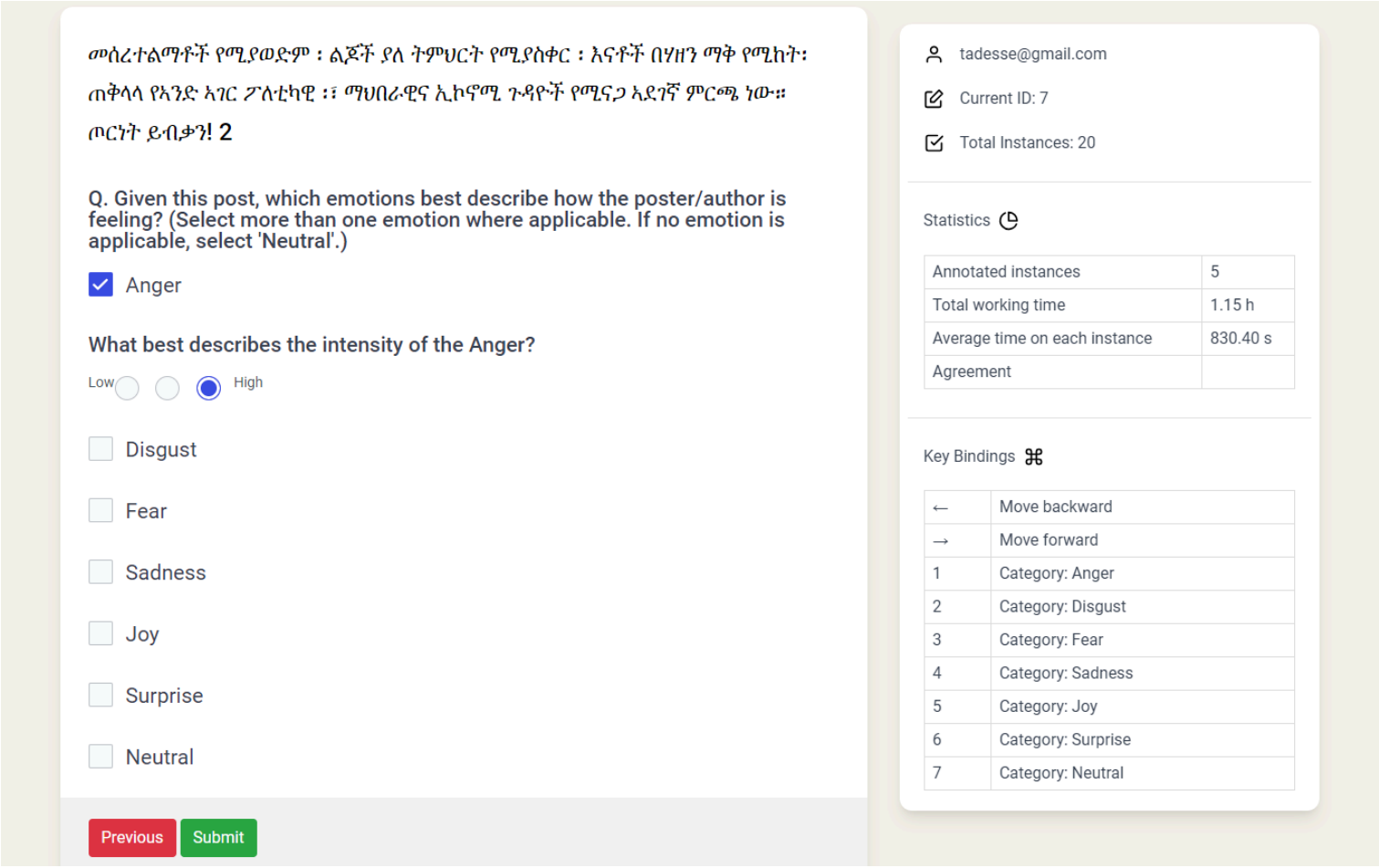
In single-label annotation, a text has either no emotion or only one of the emotions for the given list of emotions; each text is assigned one dominant emotion.

Id	Text	Emotion
sample_01	Never saw him again.	Sadness
sample_04	None of us did.	Neutral
sample_02	I love telling this story.	Joy
sample_05	I can't believe it! I won the scholarship! This is amazing!	Joy
sample_03	How stupid of him.	Anger

**9.2. Multi-label Emotion Analysis **

In multi-label emotion analysis, a text may express no emotion, a single emotion, two emotions, or multiple emotions simultaneously. A sample multi-label emotion annotation interface is shown below. The use of checkboxes allows annotators to select one or more emotion categories, along with the corresponding intensity level for each selected emotion. In multi-label emotion annotation, recording emotion intensities is common practice because different emotions may be expressed with varying degrees of strength within the

same text. Therefore, intensity annotations provide additional information beyond the mere presence or absence of an emotion.



Given annotations from multiple annotators, the final emotion labels can be determined through an aggregation process. When only binary emotion labels are available, a majority-vote approach can be used to determine the final label for each emotion category. However, when emotion intensity ratings are also collected, both the emotion labels and their intensities can be considered when making the final decision.

The following rule can be used to determine the final emotion label based on ratings from multiple annotators. The final label is binary, where 1 indicates the presence of the emotion, and 0 indicates the absence of a significant emotional expression.

For datasets annotated without considering intensities, decisions can be made by a simple majority vote. For emotion with intensity annotation, the following decision rule can be applied:

The majority vote works for simple cases. However, It may not work if you have an even number of annotators. For example, given four annotator scores (0,0,3,3): 0 (no-anger), 3 (high-anger), there is no majority vote in this case. IF (at least two annotators assign a non-

zero intensity score, e.g., 1, 2, or 3) AND (the average intensity score exceeds a predefined threshold), THEN the final emotion label = 1. ELSE the final emotion label = 0.

This approach combines annotator agreement and emotion intensity, resulting in a more reliable representation of the emotional content of the text.

Conditions for Determining the Final Emotion Label

1. At least 2 people must select 1(low), 2(medium), or 3(high): This ensures that there is a basic level of agreement among the annotators that the content contains some level of emotion (low, medium, or high).
2. Avg score > threshold: The average score given by all annotators must be greater than a predefined threshold. This ensures that the intensity of the emotion is significant enough to be considered present.

Since we have a scale from 0 to 3, a threshold of 0.5 can be a good choice. The below are example scenarios.

Example 1

- Annotator scores: 0, 0, 3,3
- Average score: (0+ 0 + 3+3) / 4 = 1.5
- Result: Since 1.5 > 0.5, final label = 1 (emotion present).

Example 2

- Annotator scores: 1, 2, 1
- Average score: (1 + 2 + 1) / 3 = 1.33
- Result: Since 1.33 > 0.5, final label = 1 (emotion present).

Example 3

- Annotator scores: 1, 1, 0
- Average score: (1 + 1 + 0) / 3 = 0.666
- Result: Since 0.66 > 0.5, final label = 1 (emotion present).

Example 4

- Annotator scores: 1, 1, 0, 0,0
- Average score: (1 + 1+ 0 + 0+0) / 5 = 0.4
- Result: Since 0.4 < 0.5, the final label =0 (no emotion present).

Example 5

- Annotator scores: 1, 2, 3, 2
- Average score = (1 + 2 + 3 + 2) / 4 = 8 / 4 = 2.0
- Result: Since 2.0 > 1.5, final label = 1 (emotion present).

This method combines both agreement among annotators and the intensity of the emotion, providing a balanced evaluation. Majority vote does not consider the intensity and may miss out on subtle emotional content. A sample format of the multi-label emotion dataset is shown below. The final binary emotion labels are then derived using the aggregation procedure described above.

Id	Text	Anger	Fear	Joy	Sadness	Surprise
sample_01	Never saw him again.	0	0	0	1	0
sample_02	I love telling this story.	0	0	1	0	0
sample_03	How stupid of him.	1	0	0	0	0
sample_04	None of us did.	0	0	0	0	0
sample_05	I can't believe it! I won the scholarship! This is amazing!	0	0	1	0	1

**9.3. Emotion Intensity **

Annotators indicate the emotions that are likely conveyed in the text and their intensity levels, i.e., low (1), medium (2) or high (3). The scores that are associated to the different intensity levels (0 [no emotion] to 3 [high emotion]) are then averaged. This average score is used to classify the emotion intensity based on its proximity to the predefined values corresponding to the intensity (i.e., low (0), 1 (low), 2 (medium), and 3 (high)).

The following is how to make majority vore during intensity annotation if the intensity labels are in likert scale 0 - 3(0 is - no emotion, 1 - low intensity, 2 - medium intensity, and 3- high intensity).

Example 1: Give an example of an emotion and a post [e.g., angry].

- Annotator scores = [1, 2, 2, 3]
- Average score = (1 + 2 + 2 + 3) / 4 = 2.0
- Intensity class = 2 (Moderate amount of emotion)

Example 2:

- Annotator scores = [0, 0, 1, 1]
- Average score = (0 + 0 + 1 + 1) / 4 = 0.5
- Intensity class = 1 (Low amount of emotion)

Example 3:

- Annotator scores = [3, 3, 3, 2]
- Average score = (3 + 3 + 3 + 2) / 4 = 2.75
- Intensity class = 3 (High amount of emotion)

Example 4:

- Annotator scores = [0, 0, 0, 1]
- Average score = (0 + 0 + 0 + 1) / 4 = 0.25
- Intensity class = 0 (No emotion)

A sample format of the multi-label emotion intensities dataset is shown below. Each emotion is assigned an intensity score ranging from 0 (not present) to 3 (high intensity).

Id	Text	Anger	Fear	Joy	Sadness	Surprise
sample_01	Never saw him again.	0	0	0	2	0
sample_02	I love telling this story.	0	0	2	0	0
sample_03	How stupid of him.	2	0	0	0	0
sample_04	None of us did.	0	0	0	0	0
sample_05	I can't believe it! I won the scholarship! This is amazing!	0	0	3	0	3

Cite this page

*Last updated on **Jun 9, 2026** by **Tadesse Destaw Belay***

Hate Speech Analysis

What is Hate Speech Analysis?

Hate speech detection is a task that focuses on identifying text that contains hate, offensive, abusive, or discriminatory language directed at individuals or groups based on attributes such as race, ethnicity, religion, gender, nationality, disability or other identity factors. The goal of this task is to automatically classify whether a piece of text expresses hate, hostility, or incitement to harm, often differentiate it from non-hateful content such as criticism, disagreement, or neutral statements. Hate speech detection is widely used in content moderation systems among various social media platforms to reduce online harassment and promote a safer digital communication ecosystem.

Researchers used various class labels, such as Muhammad et al. (2025), who used hate, abusive, normal and indeterminate as class labels, while Ayele et al. (2022, 2023) utilized hate, offensive, normal, and unsure class labels. Ayele et al. (2024) used class labels such as hate, offensive, normal and indeterminate. Davidson et al. (2017) used hate, offensive and neither class labels, while Davidson et al. (2019) used hate, and offensive class labels. Mathew et al. (2021) used hate, offensive, normal and undecided class labels.

The most agreed-upon definitions of the hate speech classes are as follows:

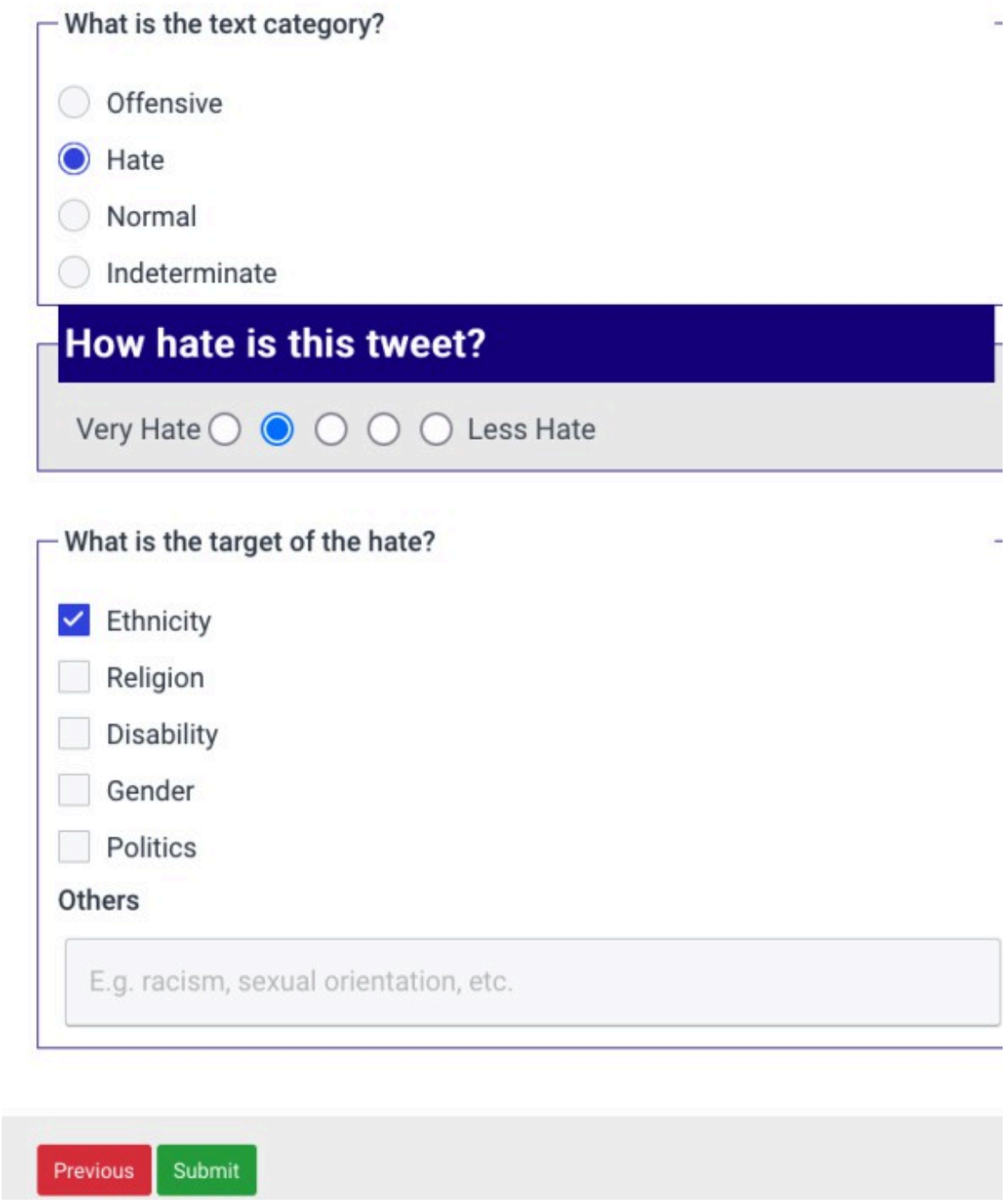
- **Offensive speech:** is any form of bad language expressions including rude, impolite, insulting, or belittling utterances intended to offend or harm an individual.
- **Hate speech:** is language content that expresses hatred towards a particular group or individual based on their group identities such as race, ethnicity, religion, gender, disability, political affiliation, or other characteristics. It also includes threats of violence associating group identities.
- **Normal **is any form of expression that does not contain any bad language belonging to any of the above classifications.
- **Indeterminate **is any tweet that is not readable or is completely written in a language other than your language of annotation.

Hate speech can target one or many targets during expression. Ayele et al (2024) presented a more general list of hate speech target labels as follows:

- **Ethnic**: if the hatred tweet targets an ethnic group identity
- **Religion**: if the hatred tweet targets a religious group identity
- **Gender**: if the hatred tweet targets a particular gender group identity
- **Disability**: if the hatred tweet targets disabled group identity
- **Politics**: if the hatred tweet targets people/entities due to political ideology
- **Unidentified target**: if the hatred tweet's target is not clearly identified/known.
- **Other**: if the hatred tweet targets other groups/group identities such as sexual orientation, racism etc.

Some studies, such as Ayele et al (2024) also examine and rate the severity of hate and offensive messages with rating scales from 1-5.

@USER @USER ሆድ ሲያውቅ ዶሮ ማታ አንቺ አጋሜ አማራ እንኳን አንቺ ጣሊያን ያውቀዋል ዘረኛ ገንጣይ አስገንጣይ የሴይጣን ልጆች።ዘመዶሽን አግር ለአግር አየተ



Cite this page

*Last updated on **Jun 9, 2026** by **Tadesse Destaw Belay***

Data Quality Control

Data quality control ensures that your sentiment labels are accurate, consistent, and reliable.

- Clear annotation guidelines: Provide precise definitions of sentiment classes (positive, negative, neutral) with examples, including tricky cases like sarcasm, negation, and mixed opinions. This reduces confusion and inconsistency.
- Multiple annotators & agreement: Assign each item to at least 2–3 annotators and measure how much they agree (e.g., using Kappa). High agreement indicates reliable labels.
- Gold-standard checks: Include a small set of pre-labeled (trusted) examples during annotation to evaluate annotator performance continuously.
- Disagreement resolution: Use majority voting or expert review to finalize labels when annotators disagree.
- Data cleaning: Remove duplicates, irrelevant content, spam, and noisy text to improve overall dataset quality.
- Class balance: Check that sentiment categories are not overly skewed (e.g., too many positives), or account for imbalance during modeling.
- Ongoing monitoring: Track annotator behavior over time to detect inconsistency or fatigue and take corrective action.

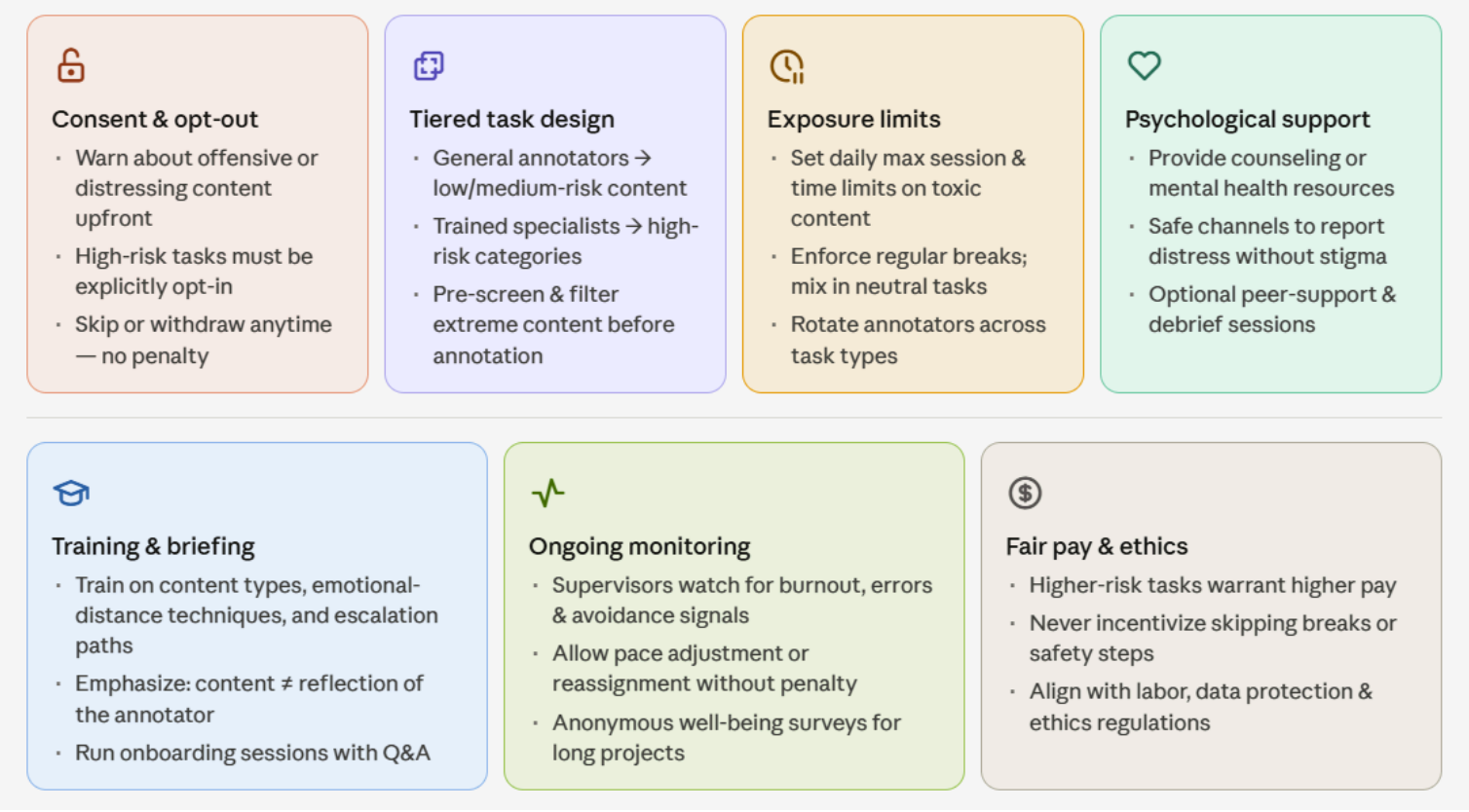
In short: combine good guidelines, multiple reviews, and continuous checks to maintain high-quality sentiment data.

*Last updated on **Jun 9, 2026** by **Tadesse Destaw Belay***

Annotator Safety and Mental Health

Protecting annotator well-being is essential, particularly when working with harmful, offensive, or emotionally distressing content. Annotators should be informed about potential risks before participation, allowed to opt out of sensitive tasks, and given the freedom to skip items or withdraw without penalty. Exposure to harmful content should be carefully managed through content filtering, workload limits, regular breaks, and task rotation. Projects should provide appropriate training, clear safety protocols, and access to psychological support resources when needed. Continuous monitoring of annotator well-being, respectful communication, protection of privacy, fair compensation, and adherence to ethical and legal standards are also critical for maintaining a safe and sustainable annotation environment.

Protecting annotators through informed consent, controlled exposure, breaks, and support systems improves both well-being and annotation quality.



Cite this page

*Last updated on **Jun 9, 2026** by **Tadesse Destaw Belay***

Glossary

A reference of terms used throughout the Playbook. Cross-references point back to the chapters where each concept is introduced in depth.

This is a starting point — additions and corrections welcome via the "Edit this page" link at the bottom.

A

Adjudication. The process of resolving disagreements between annotators, typically by a senior annotator or a designated adjudicator. Common when multiple annotators label the same item and a final "gold" label is needed. See the *Annotation Design and Workforce Management* chapter.

Annotation. Attaching structured information — labels, spans, categories, ratings — to raw data so it can be used to train or evaluate language models.

Annotation guidelines. The written specification that tells annotators exactly how to label each kind of input. Includes definitions, decision rules, worked examples, and edge cases. The single most important artifact for high inter-annotator agreement.

Annotation schema. The structural definition of what can be labeled — e.g., the set of allowed entity types in NER, or the rating scale in sentiment analysis. The schema constrains what guidelines can describe.

B

Backtranslation. Translating from the target language back to the source language to generate additional training pairs. Often used to augment low-resource translation datasets. Quality varies — verify with native speakers before training on backtranslated data.

Benchmark. A standardised dataset and evaluation protocol used to compare models. Examples relevant to African NLP: AfriSenti, NaijaSenti, AfriHate, BRIGHTER, AmhEn.

C

Cohen's kappa (κ). An inter-annotator-agreement metric for two annotators on categorical labels, corrected for chance agreement. Range: −1 to 1; conventionally κ > 0.6 is "substantial," κ > 0.8 is "almost perfect."

Consent. Documented permission from the people contributing speech, text, or images, usually including provisions on use, retention, and the right to revoke. Required for ethical and legal data work — see the *Data Collection, Curation, and Governance* chapter.

Corpus *(pl. corpora)*. A structured collection of texts, speech, or other linguistic data used for analysis or model training.

Crowdsourcing. Recruiting many distributed annotators — often online — to label data. Trade-off: scale vs. quality. Quality control techniques (gold-standard items, agreement metrics, qualification tests) become more important as crowd size grows.

D

Dataset. A curated collection of items with labels and documentation, ready to be used for training or evaluation. A dataset is a *corpus + schema + labels + documentation + license*.

Data sovereignty. The principle that data about a community belongs to that community, with associated control over storage, access, and use. Especially important for language data from indigenous and minoritised speakers.

F

Fleiss' kappa. Inter-annotator-agreement metric for more than two annotators on categorical labels — a generalisation of Cohen's kappa.

G

Gold standard. A reference labeling considered correct after adjudication or expert review. Used to evaluate annotators, evaluate models, and as the ground truth in test sets.

I

Inter-annotator agreement (IAA). A quantitative measure of how consistently different annotators produce the same labels. Low IAA suggests guidelines are unclear, the task is ambiguous, or annotators need more training.

K

Krippendorff's alpha (α). A flexible inter-annotator-agreement metric that handles missing data, multiple annotators, and different label scales (nominal, ordinal, interval, ratio).

L

License. The legal terms under which a dataset or piece of code can be used, modified, and redistributed. Common open licenses: Apache 2.0, MIT, CC-BY-SA, CC-BY-NC. Consent and license are different things — covered in the *Documentation, Data Release, and Governance* chapter.

Low-resource language. A language for which little digital data and few NLP resources exist. Most African languages fall in this category. Building useful systems requires deliberate data collection and often careful transfer from related higher-resource languages.

M

Modality. The type of input data — text, speech, image, video, or some combination. Modality-specific annotation is covered in the *Modality-Specific Task Design* chapter.

Multilingual. Covering or working across multiple languages, often with shared model parameters.

N

Named Entity Recognition (NER). Identifying spans of text that refer to named things — people, places, organisations, etc. — and labeling them with their type.

P

Parallel corpus. A corpus with the same content in two or more languages, sentence-aligned. The basis for machine-translation training.

Part-of-speech (POS) tagging. Labeling each token with its grammatical role (noun, verb, adjective, etc.).

R

Reproducibility. The property that another researcher, given the dataset, code, and reported configuration, can re-run the experiment and obtain the same results. The Playbook treats reproducibility as a first-class design goal.

S

Synthetic data. Data generated by a model rather than collected from human sources. Useful for augmentation; risky without verification because errors compound. Covered in the *LLM-Assisted and Synthetic Data Generation* chapter.

T

Tokenisation. Splitting text into the basic units a model operates on. Choices around tokenisation (subword, BPE, SentencePiece, character) materially affect downstream performance — especially in morphologically rich languages.

See also

- How to cite the Playbook
- How to contribute a chapter
- Discord community

*Last updated on **May 23, 2026** by **Shamsudddeen Hassan Muhammad***