Data Schema

> “Our annotation schema is based on the OLID multi-layer framework (Zampieri et al., 2019), which is widely used in offensive language detection benchmarks such as SemEval. We adopt this structure to ensure comparability and theoretical grounding across languages.

1. sentence_id
   - string or int
2. Lanugage
   - `ITA`
   - `KOR`
   - `SIN`
3. context
   - comment / sentence
   - the text unit on which offensiveness is judged
4. offensiveness
   - `OFF` : offensive
   - `NOT` : not offensive (including neutral or unspecified cases)

4. target_type
   - `UNT` : Untargeted
   - `IND` : Individual
   - `GRP` : Group
   - `OTH` : Other (organization, institution, event)
   - offensiveness = NOT→ `null`

5. target_group_attribute

- `Gender & Sexual Orientation`
- `Race, Ethnicity & Nationality`
- `Political Affiliation`

- `Religion`
- `Miscellaneous`
- defined only when `target_type = GRP`; otherwise `null`

6. offensive_span
   - Level A span he minimal and sufficient text span that justifies offensiveness
   - It can be tokens, word, expression..
   - that makes “context” offensive

▼ KOLD Annotation Framework (Paper-Aligned Overview)

The KOLD paper defines a **three-level hierarchical annotation framework** for offensive language.

Each level answers a different analytical question:

(1) *Is the language offensive?*

(2) *Who is being targeted?*

(3) *Which social group is being targeted?*

Level A: Offensive Language Detection

Unit of judgment: the entire comment

Annotations:

- `OFF` ∈ {OFF, NOT}

  Indicates whether the comment contains offensive language.
- `offensive_span`

  The minimal and sufficient text span that justifies *why* the comment is offensive.

The offensive span may include:

- explicit profanity,
- implicit offense such as sarcasm, metaphor, or emojis,
- multiple sentences, if offensiveness is distributed across them.

Key idea:

> Offensiveness is modeled as a binary decision plus a textual rationale (span).

Level B: Target Type Categorization

Applied only when `OFF = OFF`.

Annotations:

- `target_type` ∈ {UNT, IND, GRP, OTH}
  - **UNT (Untargeted):** general profanity without a specific target
  - **IND (Individual):** offense directed at a specific person (cyberbullying)
  - **GRP (Group):** offense targeting a social group with shared protected characteristics (hate speech)
  - **OTH (Others):** targets such as organizations, companies, or events
- `target_span`

  The span of text that explicitly or implicitly indicates the target of the offense.

The target span may overlap with the offensive span.

Key idea:

> Level B separates whether an offense is targeted and what kind of target it has.

Level C: Target Group Identification

Applied only when `target_type = GRP`.

Level C has a **two-level structure**:

1. **Target Group Attribute** (superclass / grounds of targeting)
2. **Target Group** (specific social or demographic group)

Properties:

- Multi-label annotation is allowed (one comment may target multiple groups).

- The target group attribute functions as a **superclass** of specific target groups.

Target Group Attributes defined in KOLD (5 total):

- Gender & Sexual Orientation
- Race, Ethnicity & Nationality
- Political Affiliation
- Religion
- Miscellaneous

Groups such as **Disabled** or **Feminist** are placed under **Miscellaneous**.

This is an **intentional design choice**, not an omission: these groups are treated as ideological or evaluative categories rather than primary identity attributes in the dataset design.

Key idea:

> Level C captures which social grounds are mobilized when a group is targeted.

Summary

- **Level A:** Is the language offensive, and where is the evidence?
- **Level B:** Is someone being targeted, and what kind of target is it?
- **Level C:** If a group is targeted, *which social category* does it belong to?

This hierarchical structure allows KOLD to represent offensive language not only as a binary label, but as a **structured phenomenon involving targets, social categories, and contextual justification.**