

Overview of Annotation Tasks I

Classification Tasks

Jin Zhao

Brandeis University

February 2, 2025

Today's Agenda

- ① Quick review of corpus selection
- ② Introduction to classification annotation
- ③ Sentiment analysis and opinion mining
- ④ Topic classification
- ⑤ Intent detection
- ⑥ Multi-label vs. multi-class annotation
- ⑦ Annotation challenges in classification

Semester Project: Conceptualize annotation task (written submission)

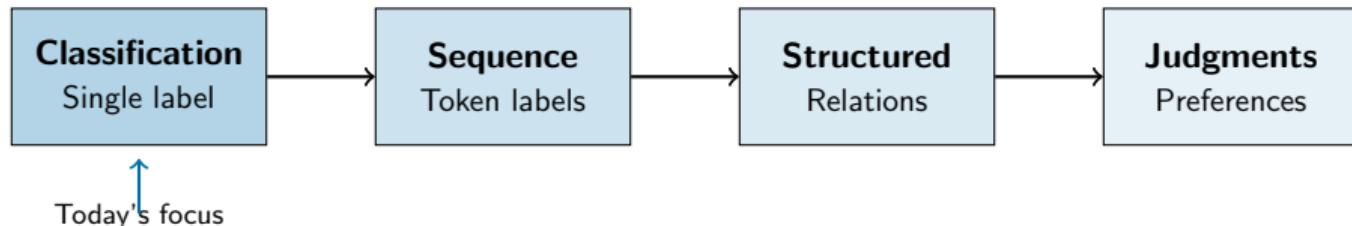
Quick Review

From last lecture on corpus selection:

- **MAMA Framework:** Medium, Annotation, Multimodality, Availability
- **Sampling strategies:** Random, stratified, targeted
- **Licensing:** CC-BY, CC-BY-NC, CC-BY-SA
- **Decontamination:** Ensuring test data isn't in LLM training sets

Key insight: Good corpus selection is the foundation for successful annotation projects.

The Task Spectrum Revisited



Classification: Assign one or more labels to a text unit

- Simplest annotation type
- Typically high inter-annotator agreement
- Most amenable to LLM annotation

What is Classification Annotation?

Definition: Assigning categorical labels to text units (documents, sentences, phrases)

Single-label (Multi-class):

- One label per instance
- Mutually exclusive categories
- Example: Positive/Negative/Neutral

Multi-label:

- Multiple labels per instance
- Non-exclusive categories
- Example: Topics (sports + politics)

Key question: What unit are you labeling? (Document? Sentence? Phrase?)

Sentiment Analysis

Goal: Determine the sentiment or opinion expressed in text

Annotation Scales

- **Binary:** Positive / Negative
- **Ternary:** Positive / Neutral / Negative
- **Likert scale:** 1-5 or 1-7 ratings
- **Continuous:** -1.0 to +1.0

Example annotations:

- “I love this movie!” → Positive
- “The service was terrible.” → Negative
- “The movie was two hours long.” → Neutral

Sentiment Annotation Challenges

Ambiguous cases:

- “Not bad” – positive?
- “Could be better” – negative?
- “Interesting” – neutral?
- Sarcasm and irony

Mixed sentiment:

- “Great food, terrible service”
- Requires aspect-level annotation

Context dependence:

- Domain-specific language
- Cultural differences
- Target of sentiment

Guideline considerations:

- What counts as neutral?
- How to handle implicit sentiment?
- Author vs. reader perspective?

Stanford Sentiment Treebank (SST)

Key innovation: Phrase-level sentiment annotation

- Built on Rotten Tomatoes movie reviews
- 11,855 sentences, 215,000 phrases
- 7-point Likert scale → 5 classes
- Each annotation by 3 Mechanical Turk workers

Insight: Sentence sentiment is compositional

- “not bad” = negation + negative → positive
- Annotate all constituents in parse tree
- No IAA calculated – averaged across 3 annotators

Lesson: Creative annotation schemes can reveal new linguistic structure

Topic Classification

Goal: Categorize documents by subject matter

Common domains:

- News categories (sports, politics, business)
- Product categories
- Academic disciplines
- Support ticket routing

Annotation decisions:

- Hierarchical vs. flat taxonomy
- Single vs. multi-label
- Level of granularity
- “Other” category handling

Challenge: Documents often span multiple topics

“The president’s economic policy affects sports team valuations”

→ Politics? Economics? Sports?

Intent Detection

Goal: Identify the user's intention in a query or utterance

Common in conversational AI

- Virtual assistants (Siri, Alexa)
- Customer service chatbots
- Task-oriented dialogue systems

Example intents:

- “Book me a flight to Boston” → book_flight
- “What’s the weather like?” → get_weather
- “Cancel my order” → cancel_order
- “Can you help me?” → request_help

Challenge: Same surface form, different intents based on context

Multi-label vs. Multi-class

Multi-class (single-label):

- Exactly one label per instance
- Labels are mutually exclusive
- Example: Sentiment (pos/neg/neu)

Annotation format:

- Radio buttons in UI
- Single categorical column

IAA implications:

- Multi-class: Standard Cohen's/Fleiss' Kappa
- Multi-label: Per-label agreement or set-based metrics

Multi-label:

- Zero or more labels per instance
- Labels can co-occur
- Example: Article topics

Annotation format:

- Checkboxes in UI
- Multiple binary columns

Designing Classification Schemas

Key principles:

- ① **Mutually exclusive** (for single-label): No instance should fit multiple categories
- ② **Exhaustive**: Every instance should fit at least one category
- ③ **Clear boundaries**: Annotators should agree on category membership
- ④ **Right granularity**: Not too broad, not too specific

Common pitfalls

- Categories that overlap (“complaint” vs. “negative feedback”)
- Missing “Other” or “None” category
- Too many fine-grained distinctions
- Labels based on world knowledge, not text

Emotion Classification: A Case Study

Task: Classify comments by expressed emotion

Possible emotion categories:

- Anger
- Appreciation
- Confusion
- Fear
- Excitement
- Joy
- Neutral
- Sadness
- Surprise
- Hate

Schema design questions:

- Single emotion or multiple emotions per comment?
- Include intensity? (mild anger vs. rage)
- What about mixed emotions?
- How to handle implicit emotions?

"It is good to have hair-splitters & lumpers." – Charles Darwin

LLM Annotation for Classification

Classification tasks are often good candidates for LLM annotation

Why LLMs work well:

- Clear, objective criteria
- Well-defined category boundaries
- Lots of similar examples in training data
- Easy to validate output format

Prompt design considerations:

- Provide clear category definitions
- Include few-shot examples for each category
- Request structured output (JSON)
- Consider confidence scores

When human annotation is still preferred:

- Highly subjective categories
- Domain-specific expertise required

Annotation Guidelines for Classification

Essential components:

- ① **Task description:** What are we classifying and why?
- ② **Category definitions:** Clear, unambiguous descriptions
- ③ **Positive examples:** Clear cases for each category
- ④ **Negative examples:** What does NOT belong
- ⑤ **Edge cases:** How to handle ambiguous instances
- ⑥ **Decision tree:** Flowchart for difficult cases

Pro tip: Start with examples, then derive guidelines

Annotate a sample, identify disagreements, then write rules that resolve them

Quality Control in Classification

Measuring agreement:

- Cohen's Kappa (2 annotators)
- Fleiss' Kappa (3+ annotators)
- Majority vote for gold standard

Common issues and solutions:

- Low agreement → Clarify guidelines, add examples
- Systematic bias → Calibration sessions
- Class imbalance → Stratified sampling
- Annotator fatigue → Shorter sessions, breaks

Target IAA: Generally aim for $\kappa > 0.7$ for classification tasks

Semester Project: Task Conceptualization

Due this week: Written submission conceptualizing your annotation task

Your submission should include:

- ① Task description:** What will you annotate and why?
- ② Task type:** Classification, sequence labeling, or other?
- ③ Proposed schema:** Initial set of labels/categories
- ④ Data source:** Where will your data come from?
- ⑤ Motivation:** Why is this task interesting/useful?

Considerations:

- Start simple – you can add complexity later
- Choose a task where you can get reasonable agreement
- Consider data availability and licensing

Next Class: Overview of Tasks II

Lecture 7 (Feb 4): Sequence Labeling Tasks

Topics:

- Named Entity Recognition (NER)
- Part-of-Speech tagging
- BIO/IOB tagging schemes
- Span annotation challenges
- Boundary decisions

Reading: Pustejovsky & Stubbs, Chapter 4

Key Takeaways

- ① **Classification** assigns categorical labels to text units
- ② **Sentiment analysis** determines opinion/attitude – watch for ambiguity and mixed sentiment
- ③ **Multi-class vs. multi-label** – choose based on whether categories are mutually exclusive
- ④ **Schema design:** Categories should be mutually exclusive, exhaustive, and clearly defined
- ⑤ **LLMs can help** with classification annotation, especially for objective categories
- ⑥ **Guidelines** should include definitions, positive/negative examples, and edge cases

Questions?

Questions?

Office Hours: Wednesdays 1-3pm, Volen 109

 jinzhaob@brandeis.edu