Data Readiness Lab
Databeredskapsverkstaden

Referensgruppmöte (2022-06-08)
Background

- October 2021 – October 2023
  Funded by VinnoVA

- Data Readiness is
  - an overlooked problem
  - often poorly understood
  - a huge bottleneck for the application of AI (Swedish Language Models)

- Project focuses on
  - the public sector
  - text data
Project Partners

- ESV
- SCB
- National Library of Sweden
- Arbetsförmedlingen
- Strängnäs Kommun
- Sveriges Kommuner och Regioner
- Polisen
- Public sector
- SME/startups
- Private sector
- Corporate/ME
- Academia
- Peltarion
- Gavagai
Challenges

Evaluations
Legal demands (e.g. GDPR)

Data
annotated

Fine-tuning (Unsupervised)
Fine-tuning (Supervised)

Applications

Legal demands (e.g. GDPR)

Fine-tuning (Supervised)
Challenges

You want to apply a language model, but it turns out the data...

- is hard to get access to
- is not in the format you expected
- does not meet your expectations w.r.t. quantity
- does not meet your expectations w.r.t. quality (noise, missing values, ...)

Data Readiness
- Availability
- Validity
- Utility
Challenges

Legal demands (e.g. GDPR)

Data

Fine-tuning (Unsupervised)

Fine-tuning (Supervised)

Applications

Evaluation

You want to apply a language model, but it turns out the data...

- Annotation quality is insufficient (=> poor model quality)
- Annotation takes too much time (= data is not enough)

Annotation

- Quality
- Efficiency
- Reusability
You want to apply a language model, but it turns out the data... contains personal information.

Anonymization
- Hands-on Application
Challenges

You want to apply a language model, but it turns out the data...

- the later you find out, the worse it is!

Evaluation
- Data Quality Assessment
- Test Data
Challenges

Legal demands (e.g. GDPR)

Evaluation
- Test Data
- Quality Assessment

Anonymization
- Hands-on Application

Data Readiness
- Availability
- Validity
- Utility

Annotation
- Quality
- Efficiency
- Reusability

Fine-tuning (Unsupervised)

Fine-tuning (Supervised)

Applications

Data

Evaluatio
n
annotated
Project Goal

“Att i stor skala möjliggöra ökad databeredskap hos behovsägare genom att tillhandahålla verktyg, ramverk och resurser.”
Project Goal

We share
• problems
• solutions

.. but not the data!
Work packages & Case Studies

**WORK PACKAGE**

- AP 1 - Tools, Frameworks, Guidelines
- AP 2 - Knowledge sharing, Education, Communication

**CASE STUDY**

- CS 1 - Strängnäs
- CS 2 - SKR
- CS 3 - SCB
- CS 4 - Polismyndigheten
- CS 5 - ESV

Data Readiness
Annotation
Anonymization
Evaluation
Data Readiness

First Step: Data Readiness Assessment

• Method: “We need to Talk About Data: The Importance of Data Readiness in Natural Language Processing” (Olsson & Sahlgren - 2021)

• Application to project partners
## Data Readiness

<table>
<thead>
<tr>
<th>Band</th>
<th>Level</th>
<th>Meaning</th>
<th>The state of data</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A-1</td>
<td><em>Utility</em></td>
<td>Ready for analysis with respect to given business objectives, hypotheses, questions, or context.</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>...</em></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>B-1</td>
<td><em>Validity</em></td>
<td>Ready to define the candidate questions and hypotheses. Includes exploratory analysis, data characterization, entity disambiguation, and de-duplication, etc.</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>...</em></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>C-1</td>
<td><em>Accessibility</em></td>
<td>Ready to be loaded in analysis software. Includes programmatic access, format conversions, and legal aspects, etc.</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>...</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C-n</td>
<td><em>Hearsay data</em></td>
<td>&quot;I'm sure AI can solve this problem. We have lots of data!&quot;</td>
</tr>
</tbody>
</table>

---

*Lawrence (2017) ['Data Readiness Levels'](Data Readiness Levels)*
Data Readiness

1: Programmatic access to data
2: Licenses in order
3: Lawful access to data
4: Ethics assessment of the data
5: Converted to suitable format
6: Characteristics known
7: Validated data
8: Stakeholders agree on business need
9: Purpose of data clear
10: Sufficient data for the use case
11: Evaluation steps clear
12: Data acquisition over time
13: Data secured
14: Risk free sharing
15: Allowed to share

Don't know  No  Partially  Yes
Data Readiness
Data Readiness

First Step: Data Readiness Assessment

• Method: “We need to Talk About Data: The Importance of Data Readiness in Natural Language Processing” (Olsson & Sahlgren - 2021)

• Application to project partners

Next Steps:

• Identification of major needs

• Creation of guidelines and tools for improved data readiness
Annotation

First Step: Audit of Tools & Methods

- Project Partners: Use Cases & Experiences
- Feature Comparison of Text Annotation Tools (Olsson, https://nlp-data-readiness.readthedocs.io)
- Job Ad Annotation at Arbetsförmedlingen (Stollenwerk, Fastlund, Nyqvist, Öhman – 2021)
Annotation

- Project Partners: Use Cases & Experiences

**Use Cases**

- Strängnäs Kommun
  - Children at Risk
  - Excel

- Arbetsförmedlingen
  - Discriminating Language
  - Label Studio
  - Doccanno + Self-made

- AI Sweden
  - Toxic Language
  - Prodigy

**Tools**

- Simplicity Search
- Bootstrapping
- Annotation Cross-checking
- Class-system adjustments

**Methods**

- Active Learning
# Feature Comparison of Text Annotation Tools

(Olsson, https://nlp-data-readiness.readthedocs.io)

<table>
<thead>
<tr>
<th>Tool</th>
<th>Prodigy</th>
<th>Label studio</th>
<th>Doccano</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tagline</td>
<td>“Radically efficient machine teaching. An annotation tool powered by active learning.”</td>
<td>“Open Source Data Labeling Tool — Simplicity built-in, no overcomplicated UIs — Supports different datatypes — Visually configurable from top to the bottom”</td>
<td>“Text annotation and data labeling Just create data and start annotating. You can be up and running in minutes.”</td>
</tr>
<tr>
<td>SaaS</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Self hosted</td>
<td>Yes, local server with web interface for annotation.</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Active Learning</td>
<td>Yes, NER, text classification, dependency parsing, pos tagging</td>
<td>No, but it can probably be implemented.</td>
<td>No, but it can probably be implemented.</td>
</tr>
<tr>
<td>Multiple annotators for same data</td>
<td>Yes</td>
<td>No, but maybe.</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Annotation

Job Ad Annotation at Arbetsförmedlingen
(Stollenwerk, Fastlund, Nyqvist, Öhman – 2021)

<table>
<thead>
<tr>
<th>METHOD</th>
<th>PURPOSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bootstrapping</td>
<td>Acceleration</td>
</tr>
<tr>
<td>Active Learning</td>
<td>Acceleration</td>
</tr>
<tr>
<td>Annotation Cross-checking</td>
<td>Quality Assurance</td>
</tr>
<tr>
<td>Class system Adjustments</td>
<td>Quality Assurance</td>
</tr>
</tbody>
</table>
Annotation

First Step: Audit of Tools & Methods
- Project Partners: Use Cases & Experiences
- Feature Comparison of Text Annotation Tools (Olsson, https://nlp-data-readiness.readthedocs.io)
- Job Ad Annotation at Arbetsförmedlingen (Stollenwerk, Fastlund, Nyqvist, Öhman – 2021)

Next Steps:
- Feature Comparison of Text Annotation Tools: Update
- Creation of Guidelines

TASK
- easy
- hard

TOOLS & METHODS
- simple
- advanced
Contact:

felix.stollenwerk@ai.se