Building Resilient Data Pipelines

A/Prof. Gianluca Demartini
Data Science Discipline
School of Information Technology and Electrical Engineering
Research Interests

• **Entity-centric Information Access** (since 2005)
  - Structured/Unstructured data (SIGIR12), Types (ISWC13, WSemJ16)
  - Entity Extraction (WWW14), Prepositions (CIKM14), Entity Cards (SIGIR19)
  - IR Eval (IRJ15, ECIR16 Best Paper, CIKM17, SIGIR18, CIKM19, WWW22, TOIS23)

• **Human-in-the-loop Information Systems** (since 2012)
  - Entity Linking (WWW12, VLDBJ), CrowdQ (CIDR 13)
  - Learnersourcing (LAK21, IEEE TLT), HITL (COMNET15, FnT17)

• **Better Crowdsourcing Platforms** (since 2013)
  - Platforms (WWW15, CSCWJ18), Experiments (CSCW21), Pricing (HCOMP14)
  - Task Allocation (WWW13, WWW16, COR), Workers (CHI15), Attacks (HCOMP18 Best Paper, JAIR), Reward (CSCW20 Hon. Mention)
  - Modus Operandi (UBICOMP17, HT19, WSDM20), Bias (SIGIR18, ECIR20 Best Paper)
  - Time (HCOMP16), Complexity (HCOMP16), Abandonment (WSDM19, TKDE)

• **Better Data** (since 2019)
  - Know. Graphs (ISWC19), Noise (WWW19), Metadata (IPM), SES (WebSci22)
  - Unknown Unknowns (ECAI20, HCOMP21), Behaviors (CIKM20)
  - Data Workers (SIGIR20, TOIS, TKDE, WWW23), Fairness (CIKM22, SIGIR23)

• **Data and AI for Public Good** (since 2020)
  - Conservation (w/ Google); Gender (w/ Wiki); Environment (ECIR21, ADCS21)
  - Fake News (w/ Meta; SIGIR20, CIKM20, IPM); Democracy (ADCS21)
The Data Science Process

- Problem Formulation
- Data Collection
- Data Preparation
- Data Analysis
- Storytelling
The Impact of Data – What’s at stake

Robust data pipelines: Enhance what humans can do with data

The data we are collecting, and the data we are not collecting
Medical doctors prescribing certain diagnostic lab tests (and not others)

The risk of uncollected data
Biased data collection because of expert decision to focus on certain aspects (and not others)

Outline

Data Collection
• Participation bias: Wikidata editors and knowledge graphs (CSCWJ + ISWC 2019)
• Unknown unknowns (HCOMP 2021)

Data Preparation and Data Quality
• Data curation behaviors (SIGIR 2020 + TOIS, TKDE)
• Behavior embeddings (CIKM 2020)

Data Labelling
• Political bias: misinformation annotations (ECIR 2020, SIGIR 2020)
• Cultural bias: socio-economic diversity of annotations (WebSci 2022)
• Bandwagon effect influencing human annotations (IP&M)

The impact of bias on ML models (CIKM 2022)
KGs and data quality

Knowledge graphs store information about entities and their relations

Data quality:
• Missing/wrong entities
• Missing/wrong information about entities
• Missing/wrong relations between entities
Wikidata - A Collaborative Knowledge Graph

KG for humans and machines
Started in 2012

Source of structured open data

Collaboratively edited

Used to power Wikipedia infoboxes
WikiData Editors

The WikiData edit history (2012-2016)

• 35M (human) edits, 8M items, 140K editors

Why do certain editors have a lifetime longer than others?

• It’s a habit: Editors with long lifespan have a constant contribution over months, while editors with short lifespan do not

• It’s not boring: Editors with a long lifespan increase the diversity of the type of their edits

Bias: longer lifespan editors contribute more and thus their views and focus dominate the KG data

Knowledge Graph - Completeness

Estimating Class Completeness
• Do we have all the cities of Germany in the KG?
Need to know class cardinality
• Easy for US States, difficult for others (need to estimate)
Estimation based on capture/recapture
• Need sampling/mentions over time

Bias: gender imbalance, that is, certain categories are more complete for male than for female

Unknown Unknowns

A Human-in-the-loop approach to UU detection:

- Extract features from human generated text
- Compare human-generated features with the features learned by the model from training images
- Identify images represented differently in the two spaces (likely to be UUs)
- Collect labels, re-train, and iterate

Figure 1: An example of UUs where the model makes a *wrong* classification but with a high confidence score. In this case, the classification model is not able to identify such mistakes automatically.

Bias in Data Preparation and Data Quality
Data Curation Behaviors


Logging Behaviors

UQCrowd Logging System

- JS code embedded in the data annotation tasks
- Send msg (for every click, keystroke, scroll, new tab, etc.) to our server

Observe human annotator online behaviors while they complete tasks

https://github.com/d-lab/uqcrowd-log
Behavior embeddings

Model human annotator behavior using embeddings
- Raw actions from logs as sequences of tokens + CBOW
- Vector representations of user behaviors

Compare user behaviors (e.g., high performers / low performers)
Changes over time
Different time granularities

<table>
<thead>
<tr>
<th>Order</th>
<th>Single Action</th>
<th>n-gram Token (n = 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ctrl+C</td>
<td>(Ctrl+C, Ctrl+V)</td>
</tr>
<tr>
<td>2</td>
<td>Ctrl+V</td>
<td>(Ctrl+V, type characters)</td>
</tr>
<tr>
<td>3</td>
<td>type characters</td>
<td>(type characters, delete characters)</td>
</tr>
<tr>
<td>4</td>
<td>delete characters</td>
<td>(delete characters, click ‘next’)</td>
</tr>
<tr>
<td>5</td>
<td>click ‘next’</td>
<td>—</td>
</tr>
</tbody>
</table>
Bias: different people use different strategies that makes them contribute more/less/differently

Datasets: https://github.com/tomhanlei/20cikm-behavior
Bias in Data Labelling
Misinformation annotation and video tagging
Crowdsourcing Truthfulness Judgements

~600 MTurk US workers

To assess truthfulness of
• US political statements (PolitiFact)
• non-US political statements (ABC)

3 scales (3, 6, and 100 levels)

Table 1: Example of statements in the PolitiFact and ABC datasets.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Speaker, Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Florida ranks first in the nation for access to free prekindergarten.”</td>
<td>Rick Scott, 2014</td>
</tr>
<tr>
<td>“Scraping the carbon tax means every household will be $550 a year better off.”</td>
<td>Tony Abbott, 2014</td>
</tr>
</tbody>
</table>
Crowd Performance VS Expert Ground Truth
Fake News labelling - Political bias

Bias: Non-expert people who vote REP are more likely to believe to statements by REP politicians.

Video of people washing hands across different socio-economic statuses

- 4 regions: Africa, Asia, Europe, the Americas; 4 different income level for each region (4*4*7=112)
- Average video duration: 13.7 seconds ($SD = 9.14$ seconds)
Bias in the annotation of SES-diverse content

- Less accurate in guessing families’ income levels for African videos.
- Videos depicting low-income households were more likely to receive negative annotations.
- Videos with higher-income families received more positive annotations.
- Negative annotations were more prevalent for videos shot in Africa than in Asia.
- Video from higher income groups more appropriate for inclusion in search results and public service announcements.

Influencing human annotators

- Presenting metadata in tasks can significantly improve the efficiency of annotations.

- Human metadata is a popular resource to assess relevance. **Strong bandwagon effect**

- The role of metadata is subject to its quality.

What happens when we train ML models with biased labels?

Demo at: https://recant.cyens.org.cy/

Periklis Perikleous, Andreas Kafkalias, Zenonas Theodosiou, Pınar Barlas, Evgenia Christoforou, Jahna Otterbacher, Gianluca Demartini, and Andreas Lanitis. **How Does the Crowd Impact the Model? A tool for raising awareness of social bias in crowdsourced training data.** In: The 31st ACM International Conference on Information and Knowledge Management (CIKM 2022) - Demo track. Atlanta, Georgia, USA, October 2022.
1. Input image:
Click here to change the image
Current image: CFD-BF-003-003-N

2. Classification task:
Select a classification task.
- Gender
- Race
- Attractiveness
- Trustworthiness

The models try to predict the depicted person's Trustworthiness.
Nine different models were trained on the same images for each task, with different (sub)sets of crowd-worker annotations. The same input image (above) was passed through each of the nine models, resulting in the following outputs (possible outputs: Low, Medium, High):

<table>
<thead>
<tr>
<th>Model</th>
<th>Model Description</th>
<th>Classification Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFD Annotators</td>
<td>Model trained on the norming data provided with the CFD.</td>
<td>High</td>
</tr>
<tr>
<td>All Annotators</td>
<td>Model trained using all the annotations for all images.</td>
<td>Medium</td>
</tr>
<tr>
<td>Random</td>
<td>Model that simulates the case where annotators generate labels without considering the image content.</td>
<td>Medium</td>
</tr>
<tr>
<td>Men</td>
<td>Model trained using all the annotations provided by male crowdworkers.</td>
<td>Low</td>
</tr>
<tr>
<td>Women</td>
<td>Model trained using all the annotations provided by female crowdworkers.</td>
<td>Medium</td>
</tr>
<tr>
<td>Black</td>
<td>Model trained using all the annotations provided by Black crowdworkers.</td>
<td>Medium</td>
</tr>
<tr>
<td>Asian</td>
<td>Model trained using all the annotations provided by Asian crowdworkers.</td>
<td>Low</td>
</tr>
<tr>
<td>White</td>
<td>Model trained using all the annotations provided by White crowdworkers.</td>
<td>Medium</td>
</tr>
<tr>
<td>Latino</td>
<td>Model trained using all the annotations provided by Latino crowdworkers.</td>
<td>High</td>
</tr>
</tbody>
</table>
Lessons learned and what to do

- Bias is present in human-generated data is propagated in data pipelines
- Bias comes from human annotators as much as system design choices

- We need to track and profile data bias across the data pipeline
- Select and diversify the sources of the labels (i.e., human annotators)
- **Bias management** instead of bias removal
  - Demartini et al. “Data Bias Management”, in *Communications of the ACM*  
    https://arxiv.org/abs/2305.09686