Towards Conversational Data Annotation: Personalized Annotation Explanation Generation via Large Language Models

Thesis Proposal

Mark Nagengast Porro

Contents

Motivation

- Data annotation often done by multiple annotators

Motivation

- Data annotation often done by multiple annotators
  - diverse sociodemographic backgrounds
  - different perspectives on data

Motivation

- Data annotation often done by multiple annotators
  - diverse sociodemographic backgrounds
  - different perspectives on data
- Disagreements commonplace, due to

Motivation

- Data annotation often done by multiple annotators
  - diverse sociodemographic backgrounds
  - different perspectives on data
- Disagreements commonplace, due to
  - subjectivity; differences in annotators' characteristics
  - ambiguity; missing context, insufficient information
  - difficulty; varying expertise

Motivation

- Data annotation often done by multiple annotators
  - diverse sociodemographic backgrounds
  - different perspectives on data
- Disagreements commonplace, due to
  - subjectivity; differences in annotators' characteristics
  - ambiguity; missing context, insufficient information
  - difficulty; varying expertise
- Disagreement often filtered out

Motivation

- Data annotation often done by multiple annotators
  - diverse sociodemographic backgrounds
  - different perspectives on data
- Disagreements commonplace, due to
  - subjectivity; differences in annotators' characteristics
  - ambiguity; missing context, insufficient information
  - difficulty; varying expertise
- Disagreement often filtered out
  - may be resolvable

Motivation

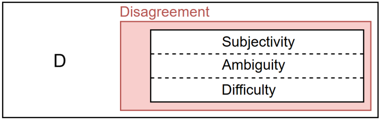
- Data annotation often done by multiple annotators
  - diverse sociodemographic backgrounds
  - different perspectives on data
- Disagreements commonplace, due to
  - subjectivity; differences in annotators' characteristics
  - ambiguity; missing context, insufficient information
  - difficulty; varying expertise
- Disagreement often filtered out
  - may be resolvable
- Provide annotators with explanations that

Motivation

- Data annotation often done by multiple annotators
  - diverse sociodemographic backgrounds
  - different perspectives on data
- Disagreements commonplace, due to
  - subjectivity; differences in annotators' characteristics
  - ambiguity; missing context, insufficient information
  - difficulty; varying expertise
- Disagreement often filtered out
  - may be resolvable
- Provide annotators with explanations that
  - address subjectiveness (objectify)
  - elaborate on issue (disambiguate, simplify)

Problem

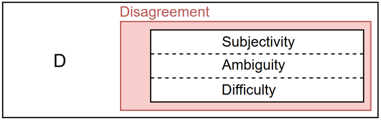
Overview



- Dataset $\mathcal{D}$ contains subset of instances with dissimilar annotations

Problem

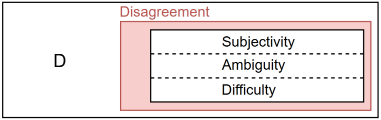
Overview



- Dataset $\mathcal{D}$ contains subset of instances with dissimilar annotations
- Generate explanations for all possible labels

Problem

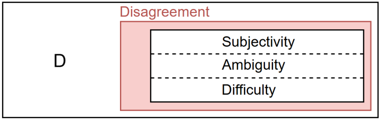
Overview



- Dataset $\mathcal{D}$ contains subset of instances with dissimilar annotations
- Generate explanations for all possible labels
  - also including sociodemograhic backgrounds, other metadata

Problem

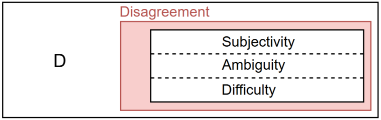
Overview



- Dataset $\mathcal{D}$ contains subset of instances with dissimilar annotations
- Generate explanations for all possible labels
  - also including sociodemograhic backgrounds, other metadata
- Reannotate $\mathcal{D}$ with the addition of the explanations

Problem

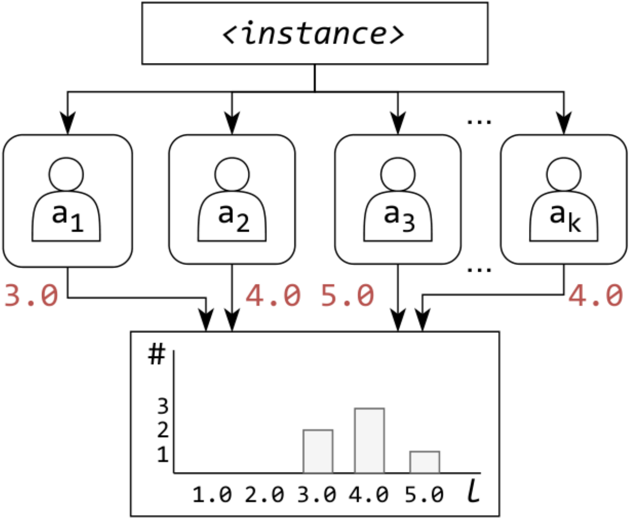
Overview



- Dataset $\mathcal{D}$ contains subset of instances with dissimilar annotations
- Generate explanations for all possible labels
  - also including sociodemograhic backgrounds, other metadata
- Reannotate $\mathcal{D}$ with the addition of the explanations
- Quantify helpfulness based on change in agreement

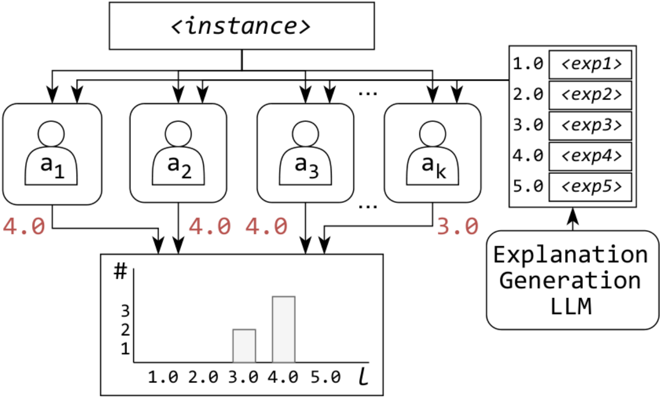
Problem

Default



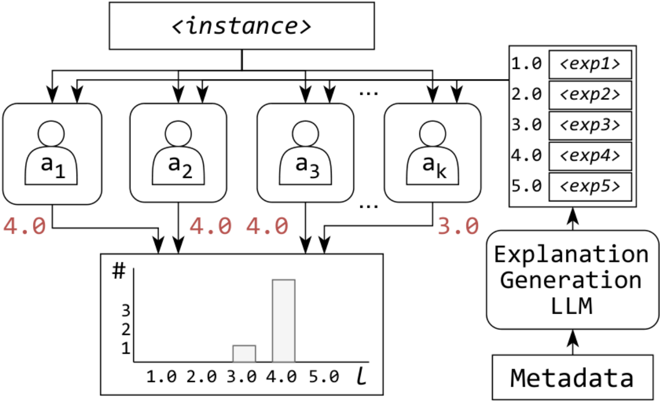
Problem

Basic Explanations



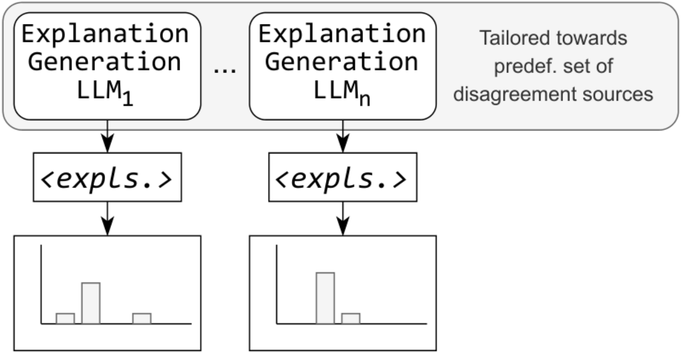
Problem

Inclusion of Metadata



Problem

Tailored Models



Research Questions

- How does providing explanations for each labeling option in a data annotation task affect the agreement between annotators?
- How can metadata such as sociodemographic information about annotators and existing annotations be used to improve explanation generation?

Thesis Goals and Tasks to Tackle Each Goal

- Resolve/mitigate disagreement in data annotation using explanations in natural language for all annotation choices

Thesis Goals and Tasks to Tackle Each Goal

- Resolve/mitigate disagreement in data annotation using explanations in natural language for all annotation choices
- Develop LLMs capable of generating explanations

Thesis Goals and Tasks to Tackle Each Goal

- Resolve/mitigate disagreement in data annotation using explanations in natural language for all annotation choices
- Develop LLMs capable of generating explanations
  - basic explanations
  - inclusion of metadata
  - tailored towards certain sources of disagreement

Thesis Goals and Tasks to Tackle Each Goal

- Resolve/mitigate disagreement in data annotation using explanations in natural language for all annotation choices
- Develop LLMs capable of generating explanations
  - basic explanations
  - inclusion of metadata
  - tailored towards certain sources of disagreement
- Multiple runs of annotations

Thesis Goals and Tasks to Tackle Each Goal

- Resolve/mitigate disagreement in data annotation using explanations in natural language for all annotation choices
- Develop LLMs capable of generating explanations
  - basic explanations
  - inclusion of metadata
  - tailored towards certain sources of disagreement
- Multiple runs of annotations
- Analyze how annotations change as a result

Thesis Goals and Tasks to Tackle Each Goal

- POPQUORN (Pei & Jurgens, 2023)

Task	Description	Data	Total Annotations	Number of Annotators	Instances	Average Labels per Instance
Offensiveness rating	Rate comment offensiveness using a 1-5 scale	Ruddit	13,036	262	1,500	8.7
Question Answering	Read a passage and answer a question through highlighting the text	SQuAD	4,576	459	1,000	4.6
Text rewriting / Style transfer	Read an email and revise it to make it sound more polite	Enron	2,346	257	1,429	1.6
Politeness Rating	Rate the politeness of an email using a 1-5 scale	Enron	25,042	506	3,718	6.7
POPQUORN			45,000	1,484	7,647	–

Thesis Goals and Tasks to Tackle Each Goal

- POPQUORN (Pei & Jurgens, 2023)

Task	Description	Data	Total Annotations	Number of Annotators	Instances	Average Labels per Instance
Offensiveness rating	Rate comment offensiveness using a 1-5 scale	Ruddit	13,036	262	1,500	8.7
Question Answering	Read a passage and answer a question through highlighting the text	SQuAD	4,576	459	1,000	4.6
Text rewriting / Style transfer	Read an email and revise it to make it sound more polite	Enron	2,346	257	1,429	1.6
Politeness Rating	Rate the politeness of an email using a 1-5 scale	Enron	25,042	506	3,718	6.7
POPQUORN			45,000	1,484	7,647	–

  - Given annotators' backgrounds (gender, race, age, occupation, education)

Thesis Goals and Tasks to Tackle Each Goal

- POPQUORN (Pei & Jurgens, 2023)

Task	Description	Data	Total Annotations	Number of Annotators	Instances	Average Labels per Instance
Offensiveness rating	Rate comment offensiveness using a 1-5 scale	Ruddit	13,036	262	1,500	8.7
Question Answering	Read a passage and answer a question through highlighting the text	SQuAD	4,576	459	1,000	4.6
Text rewriting / Style transfer	Read an email and revise it to make it sound more polite	Enron	2,346	257	1,429	1.6
Politeness Rating	Rate the politeness of an email using a 1-5 scale	Enron	25,042	506	3,718	6.7
POPQUORN			45,000	1,484	7,647	–

  - Given annotators' backgrounds (gender, race, age, occupation, education)
  - Example:
    `635, 0, Don't tell them just let them and their liniage die out so we can be free of humans without brain cells, 3.0, Man, White, 35-39, Unemployed, High school diploma or equivalent`

Outline

- Introduction
- Related work
    - Human-LLM collaboration
        - Data annotation
    - Annotator Disagreement
    - Explanation generation
    - LLM personalization
- Datasets
    - POPQUORN
- Approach
    - Sources of disagreement
    - Fine-tuning LLMs
        - Predicting disagreement source
        - Generating tailored explanations
- Annotation Experiments
- Evaluation
    - Inter-annotator agreement
    - Impact of disagreement type
- Conclusion