ADS Grant Roadmap - Version 1

16 January 2024

Dataset

1. EXIST: `http://nlp.uned.es/exist2023/`
2. SemEval 2023 - Task 10: `https://codalab.lisn.upsaclay.fr/competitions/7124`
3. SemEval 2023 - Task 11: `https://codalab.lisn.upsaclay.fr/competitions/6146`

Table 1: Strengths and Weaknesses of Datasets

Dataset	**Strengths**	**Weaknesses**
EXIST	Multi annotators (data available)	Not clear structure (we can ask)
	Other information about annotators (Gender/Age)	
	Different levels of Sexism detection	
	In Spanish and English	
SemEval 2023 - Task 10	Multi annotators (data not available)	Data of annotators is not public (we can ask)
	Clear structure and document (available)	
SemEval 2023 - Task 11		

Proposed Model Structure

- The project will employ transformer models, such as BERT, for prediction.
- SHAP (SHapley Additive exPlanations) will be used to identify influential tokens or phrases for generating explanations.
- An A/B testing framework will be established to evaluate the impact of model explanations on annotator agreement. **[Consultant]**

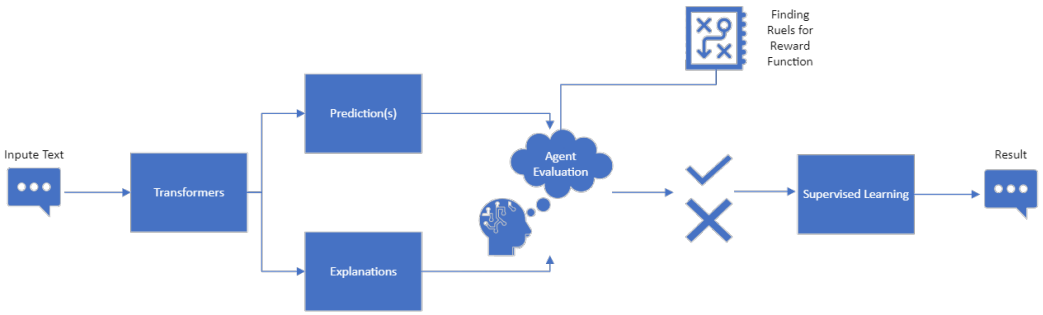

Figure 1: Methodology

Phases

1. **Data Preparation (Months 1-2):**
   - Access and preprocess the selected datasets (EXIST, SemEval Task 10/11). [**Assistant**]
   - Access and read the structure of Annotations. [**Assistant**]
   - Work on the model structure that provides both prediction and explanation.
2. **Annotation and Calibration (Months 3-4):**
   - Make a survey structure suitable for our model. [**Assistant/Consultant**]
   - Find a proper platform and create the survey. [**Assistant**]
   - Annotate a subset of data with multiple annotators.
   - Implement a structured training and calibration process for annotators.
3. **Model Integration and Prediction (Months 5-6):**
   - Utilize transformer models (e.g., BERT) for prediction on annotated data.
   - Generate explanations using SHAP for model predictions.
4. **Annotator Agreement Analysis (Months 7-8):**
   - Calculate Inter-rater Reliability (IRR) metrics (e.g., Cohen's Kappa) for annotator consensus.
   - Analyze confusion matrices to identify agreement patterns.
5. **A/B Testing (Months 7-8):**
   - Conduct A/B testing with two groups of annotators: one with model predictions and another with both predictions and explanations.
6. **Feedback Mechanism (Months 9-10):**
   - Implement a feedback mechanism for annotators to report ambiguous or unclear predictions and explanations.
   - Assess systematic bias and sensitivity to explanation types.
7. **Data Analysis and Reporting (Months 9-10):**
   - Perform statistical analysis of annotator responses. [**Assistant**]
   - Examine the impact of explanations on annotator agreement.
   - Prepare a research paper and final report.

Metrics

- Metrics include Inter-rater Reliability (IRR) metrics (e.g., Cohen's Kappa, Fleiss' Kappa, Krippendorff's Unitizing Alpha) to measure annotator consensus.
- A confusion matrix will be used to identify agreement patterns, especially False-Positive and False-Negative cases.

Output

1. Quantitative measure of annotator agreement.
2. Research paper reporting the influence of model explanations on annotator agreement.
3. Feedback analysis to improve model predictions and explanations.

Next steps

1. Working on assessing the dataset and annotation structure to find the best dataset for our project.
2. scheduling monthly update meetings.