Challenge AI's Mind: A Crowd System for Proactive AI Testing

Siwei Fu, Xiaotong Liu, Anbang Xu, Rama Akkiraju

Presenter: Dr. Siwei Fu
fusiwei339@gmail.com
Zhejiang Lab

About the Presenter:

Dr. Siwei Fu is an associate research scientist in Zhejiang Lab. His main research interests include: visual analytics, intelligent user interface, and natual language interface. He received his Ph.D. degree in Computer Science and Engineering from the Hong Kong University of Science and Technology.

Background

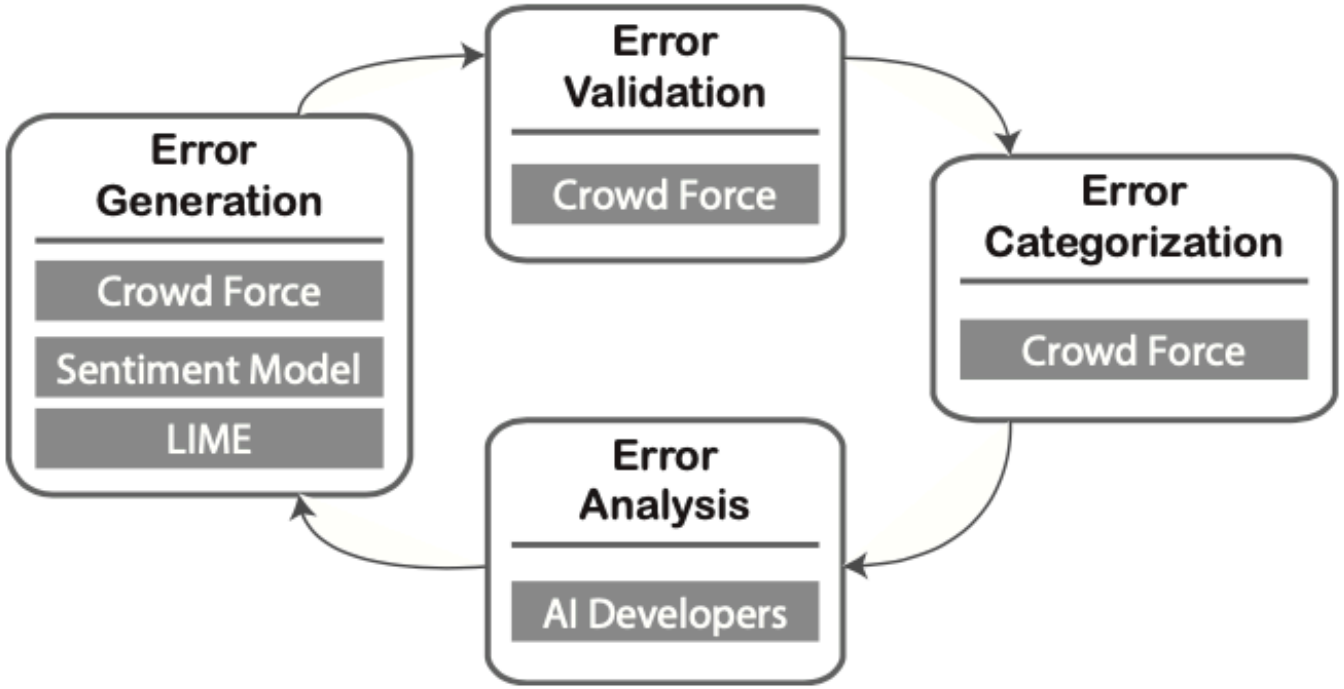
- Artificial intelligence (AI) become a technology renaissance and is beginning to solve problems in many domains.
- It performs well under **single-score metrics** such as precision and recall
- However, AI applications can fail in **critical and embarrassing cases**
  - Recent AI-powered facial recognition systems of Microsoft, IBM, and Face++ have 34% more errors with dark-skinned females than light-skinned males

Introduction

- We propose ***proactive testing***, a novel approach that evaluates the performance of AI techniques with dynamic and well-crafted dataset collected using crowd intelligence
  - It **extends the coverage** of the testing dataset by dynamically collecting external dataset.
  - AI developers are allowed to query additional dataset **belonging to certain categories** to target corner cases
- Proactive testing is an approach to **discovering unknown error and bias** of a model, and providing **a comprehensive evaluation** of the model's performance regarding all test cases.

Introduction

- We contribute a hybrid system, Challenge.AI, that combines **human intelligence** and **machine learning techniques** to assist AI developers in the process of proactive testing.



Formative Study

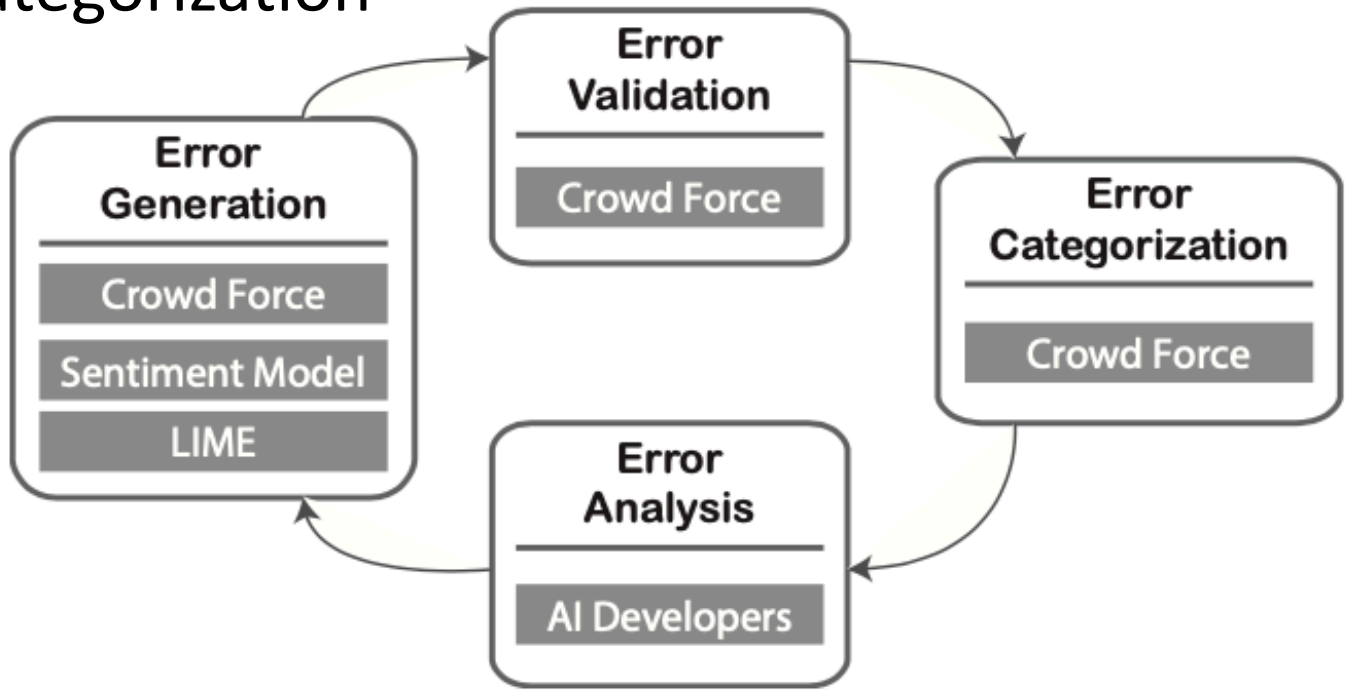
- Goal: understand current practice of model testing, the challenges faced by AI developers, and potential opportunities of our system.
- Interview five AI developers in an IT company
- Focusing on sentiment analysis models
- 30 minutes for each interview
  - Past experience in sentiment analysis
  - Observation
  - Challenges they encounter
- Results: four requirements to guide the design of Challenge.AI

Formative Study

- R1: Error generation:
  - allow AI developers to collect corpus of certain category to thoroughly test the performance of models
- R2: Error validation:
  - borrow the crowd to manually validate the sentiment of each generated sample
- R3: Error categorization:
  - validate the category of samples generated by the crowd
- R4: Error analysis:
  - Analyze mis-classified samples would reveal insights to the model

Challenge.AI

- Explanation-based error generation
- Accountability via machine learning
- Error validation and categorization
- Error analysis



Explanation-based error generation

Instruction

ⓐ

1. Introduction

Sentiment analysis is classifying whether the expressed opinion in a sentence is positive, negative, or neutral.

2. Requirements

a. Please write sentences with**"Mixed Sentiment"** containing both positive and negative sentiment indicators.

b. Please create sentences that can fail AI in sentiment analysis.

3. Examples

Utterance	Validated Sentiment	AI Mis-classified It As
Although he was having lots of bad luck, he was still positive with his life.	Positive	Neutral
It was hard going to school in a Maori dominated school in New Zealand, but I finally made it.		Negative
My stomach feel so much better from yesterday's pain		
He though I was sad but I was happy		
The day is sad, but that does not take away my happiness		

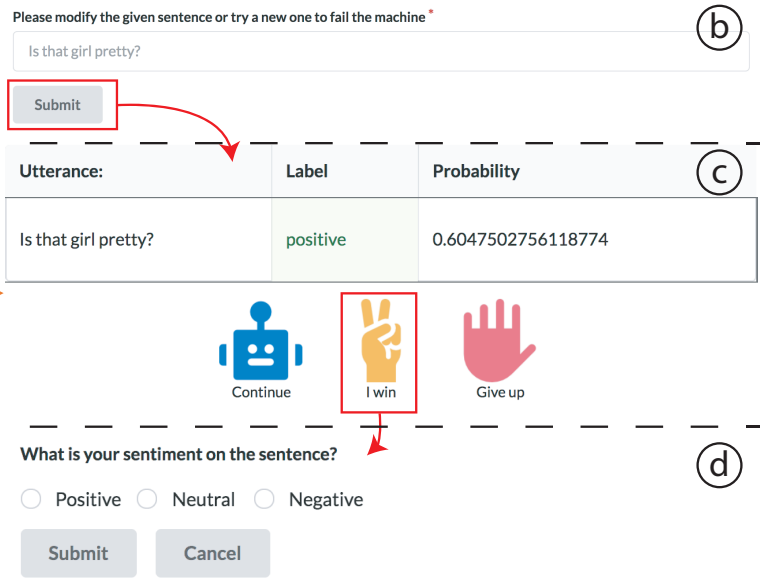
4. Notes

a. You won't be paid if you copy and paste the sentences in the examples. Please write sentences that are totally different from examples.

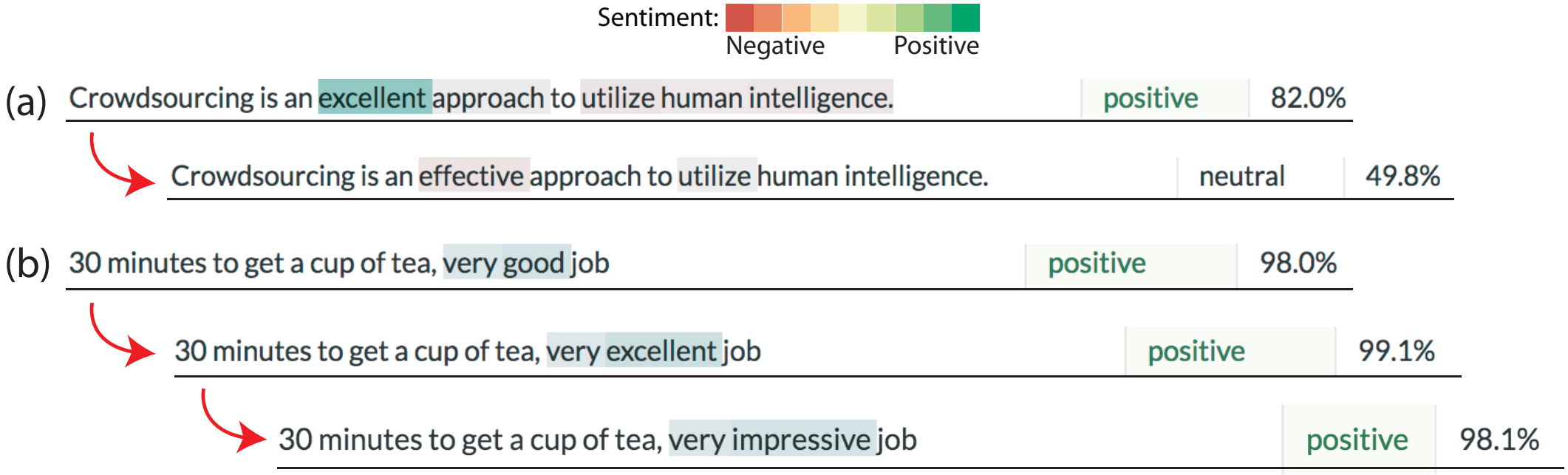
b. Please DO NOT write duplicate input and non-English input. Otherwise you won't be paid.

c. If your sentences successfully fail the AI after the validation, you will get **5X BONUS** for each sentence. At the same time, if your sentences belong to the category "Mixed sentiment", you will get**10X BONUS**.

d. If your performance is bad, you won't be able to participate the game.

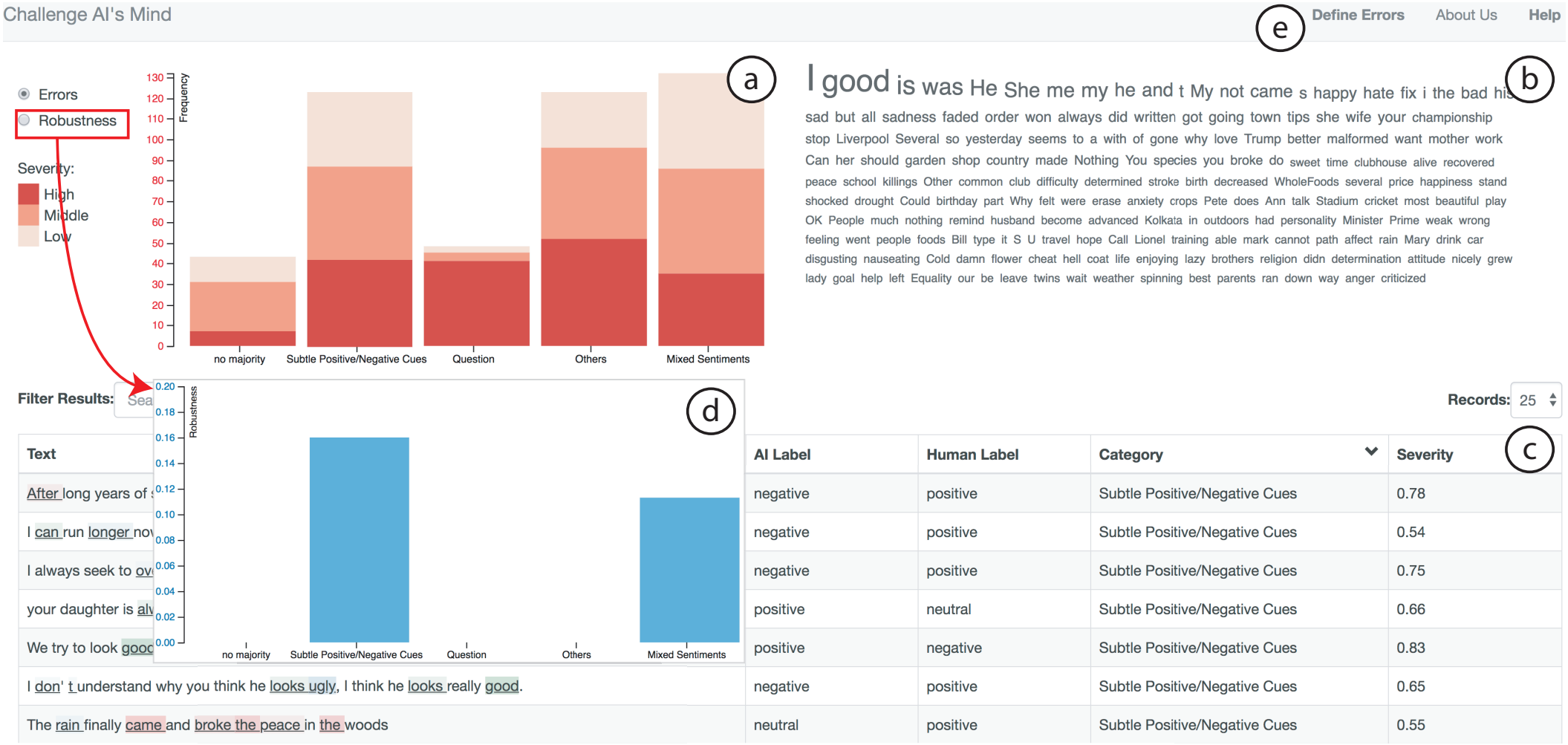


Accountability via machine learning



The usage of LIME in two cases.
(a) shows how LIME helps crowd workers modify the input sentence to successfully fool the analyzer.
(b) demonstrates how LIME facilitates workers to continuously generate adversarial samples.

Error analysis



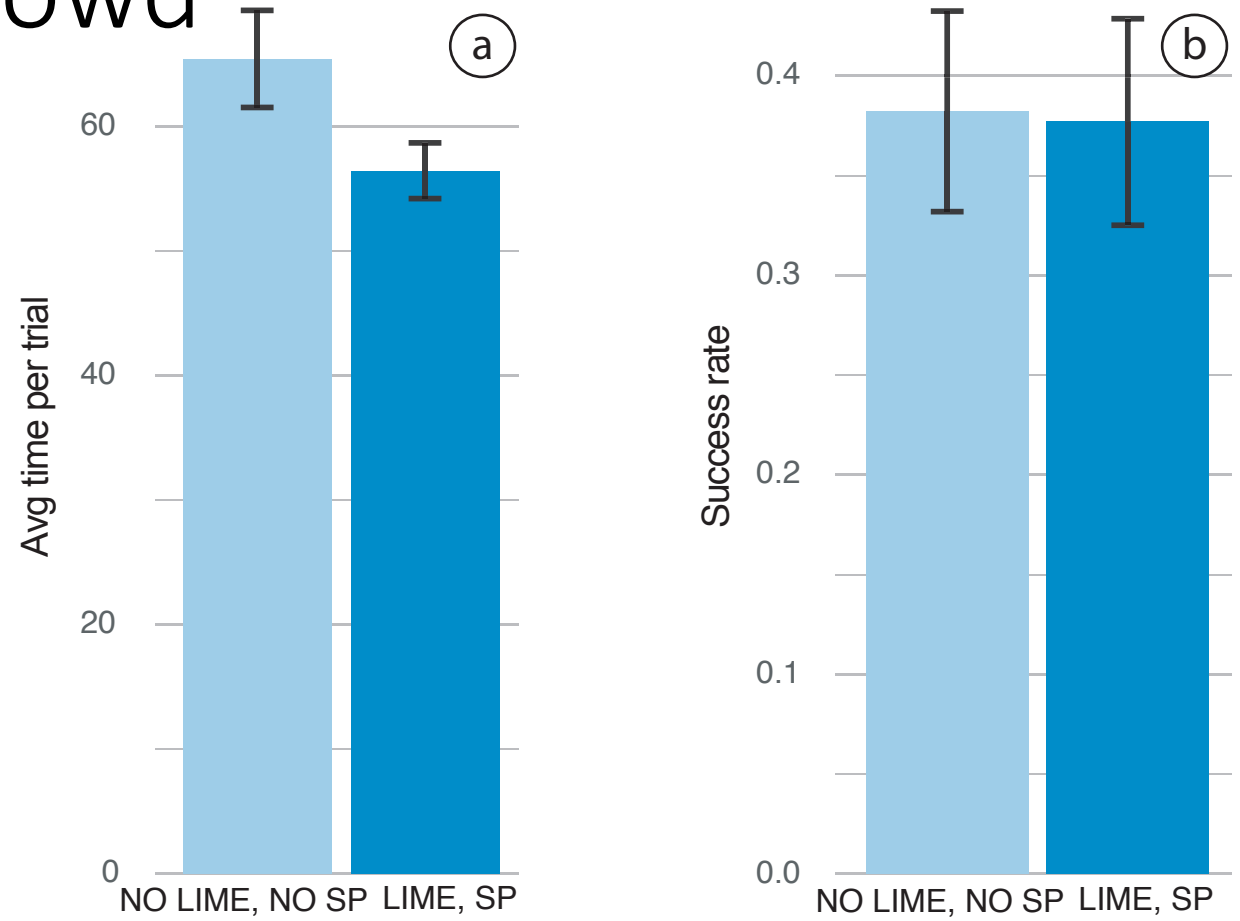
Evaluation with the Crowd

- Goal: investigate how **different prompts** in error generation affect the performance of the crowd in crafting errors.
- Construct prompts based on different combination of accountability (LIME) and starting points (SP)
- Between-subject
- Two conditions:
  - NO LIME & NO SP
  - LIME & SP
- Metrics:
  - Average time per trial
  - Success rate

Evaluation with the Crowd

	LIME, SP	No LIME, No SP	Total
N_{total}	262	293	555
N_{valid}	75	108	183
#workers	66	46	112

Statistics of error generation based on two prompt conditions



(a) shows the bar chart displaying average time per trial for each worker under two conditions.
(b) shows how crowd workers differ in success rate. The error bars demonstrate standard errors.

Evaluation with AI developers

- Process:
  - First session: obtain initial categorization for errors
  - Running Challenge.AI
    - Generate errors belonging to these categories
    - Conducted validation and categorization for crafted sentences
  - Second session: understand the usefulness and limitations of Challenge.AI from the perspective of AI developers

First Session

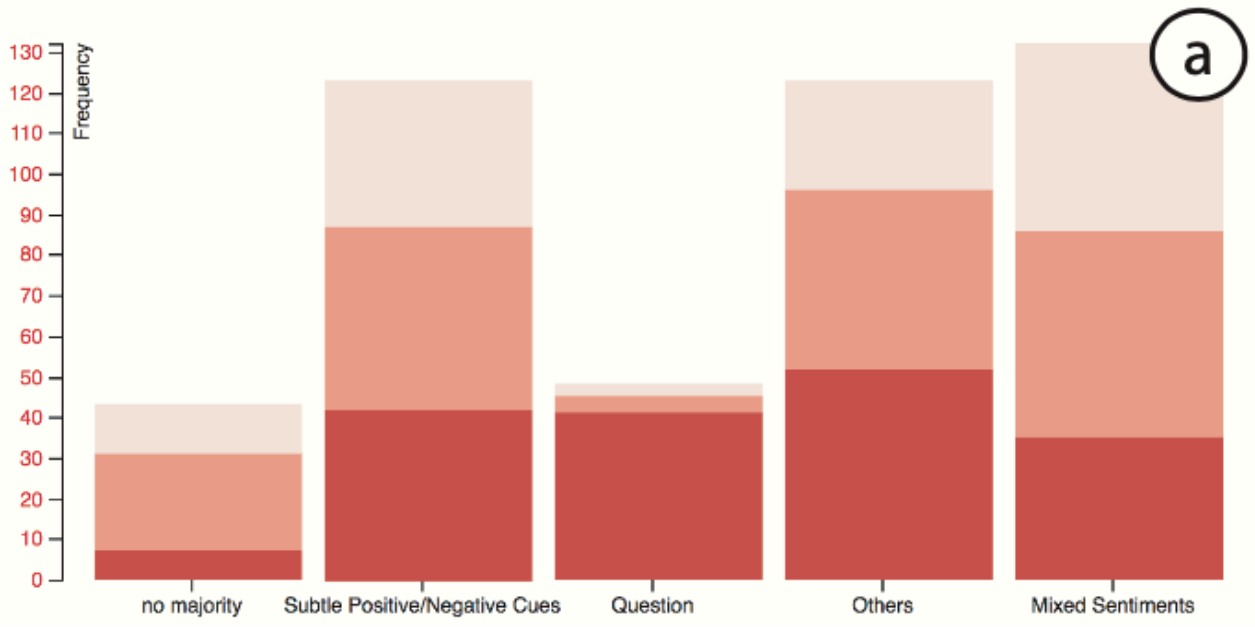
- **Subtle Sentiment Cues**
  - a sentence is either positive or negative, and has positive or negative indications
- **Mixed-sentiment**
  - refers to sentences containing both positive cues and negative indicators
- Questions
  - Sentences with a question mark
- Others
  - More general

Running Challenge.AI

- Go through three main components of Challenge.AI, e.g., error generation, validation, and categorization.
- Focused on the two categories, i.e., “Subtle Sentiment Cues” and “Mixed- sentiment”
- Finally, we obtained **555 samples** that **112 crowd workers** generated to have successfully failed the model, where **23** errors are categorized as “Subtle Sentiment Cues” and **44** are “Mixed-sentiment”

Second Session

- Getting a gist



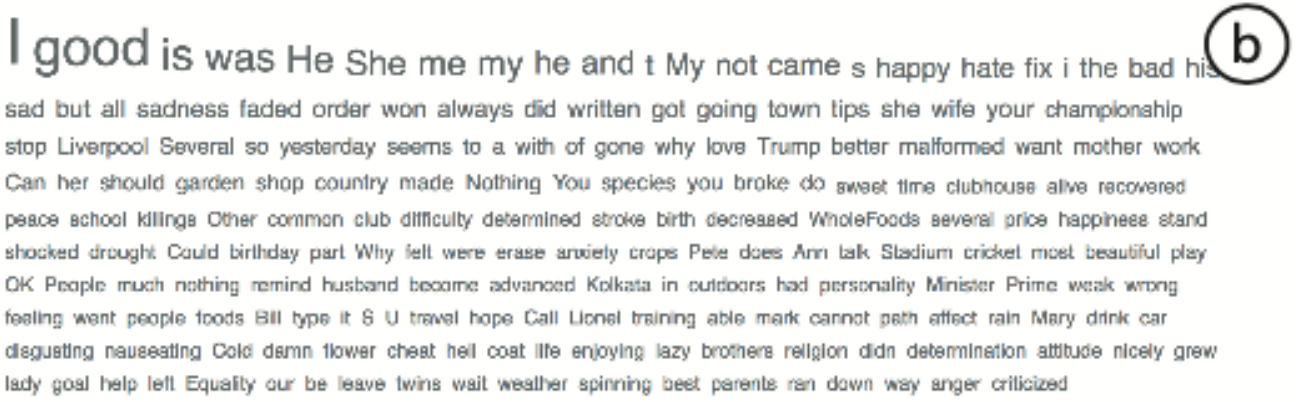
The samples belonging to “Question” attracted participants’ attention because high-severity errors account for the majority in this category

Story:

If a model a has high probability to make severe errors for question sentences, we may specify a feature in feature engineering to detect whether a sentence is a question or a statement. So with this feature, hopefully could help the model make decisions

Second Session

- *Examining errors by words*



Our participant first clicked “She” and the Table View updated. The participant noticed that the word contributes a lot to neutral sentences, and contributes once for negative and positive, respectively. Similarly, the participant further examined sentences containing the word “He”, and noticed that four out of eight are negative, and “He” contributes to the negative sentiment.

Story:

Well, it is interesting to see the difference between ‘She’ and ‘He’. I guess the model tends to regard ‘He’ as a negative word.” He added, *“I think that it is necessary to examine the training data (of the model) to see whether the stop words are equal in distribution for each sentiment*

Design Implications

- Include all the generated data by the crowd including those that can fail the model and those cannot
- Apply better explanation techniques
- Enhance the generation component for word-level categories
- Provide real-time feedback for proactive testing.
- Augment error analysis with advanced analytical methods.

Thank you!

Siwei Fu

fusiwei339@gmail.com

Zhejiang Lab