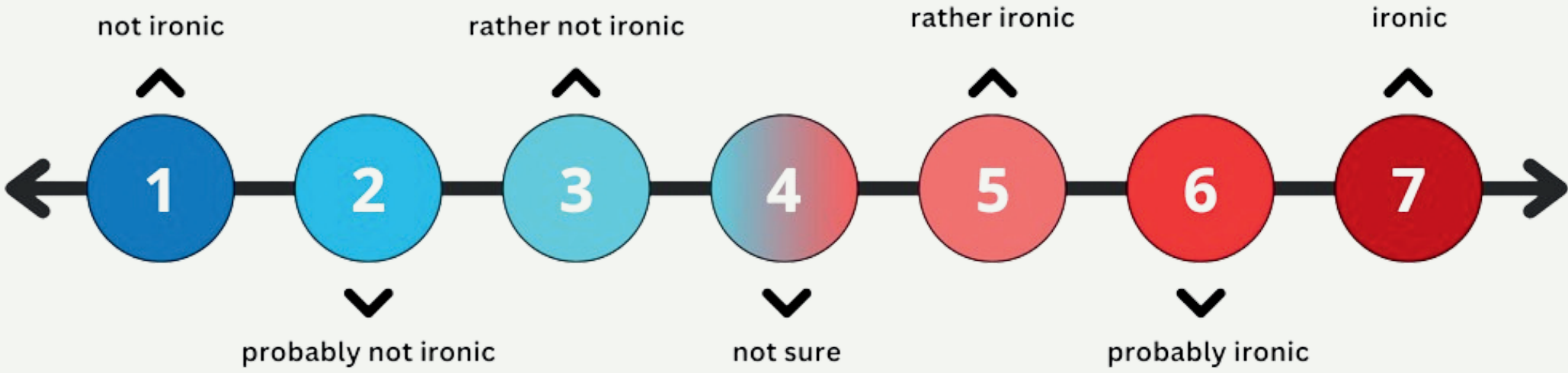


# Human and System Perspectives on the Expression of Irony: an Analysis of Likelihood Labels and Rationales

Aaron Maladry, Alessandra Teresa Cignarella,  
Cynthia van Hee, Els Lefever and Véronique Hoste

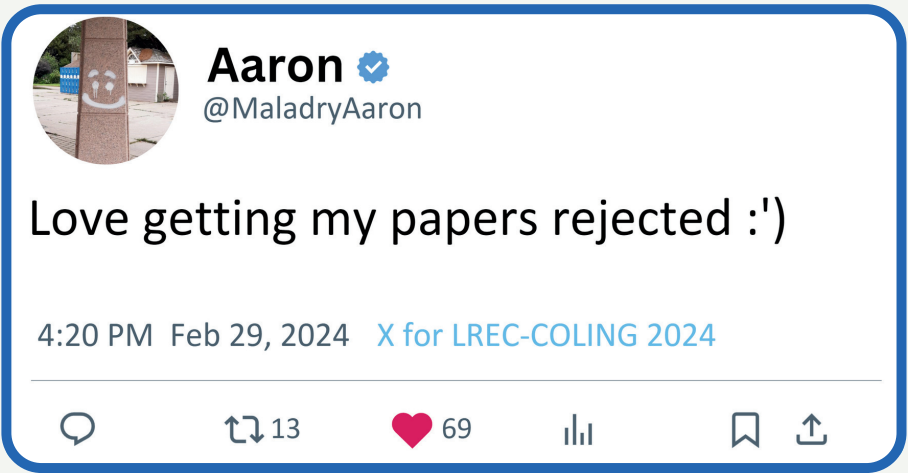
## Annotation Novelty

- irony likelihood
  - including confidence measure
- trigger words
  - contrasting elements
  - implicit knowledge



## Example:

- likelihood: 6/7
- triggers:
  - "love"
  - "papers rejected"
  - ":)"



## Explainability :

- explainability metrics:
  - sub-token attributions
  - Layer Integrated Gradients
- system importance  
VS human trigger words

	love	getting	my	papers	rejected	:)"
HUM.	1	0	0	1	1	1
INIT.	.66	-.06	.09	.36	.38	.49
NORM.	.33	0	.05	.18	.19	.25



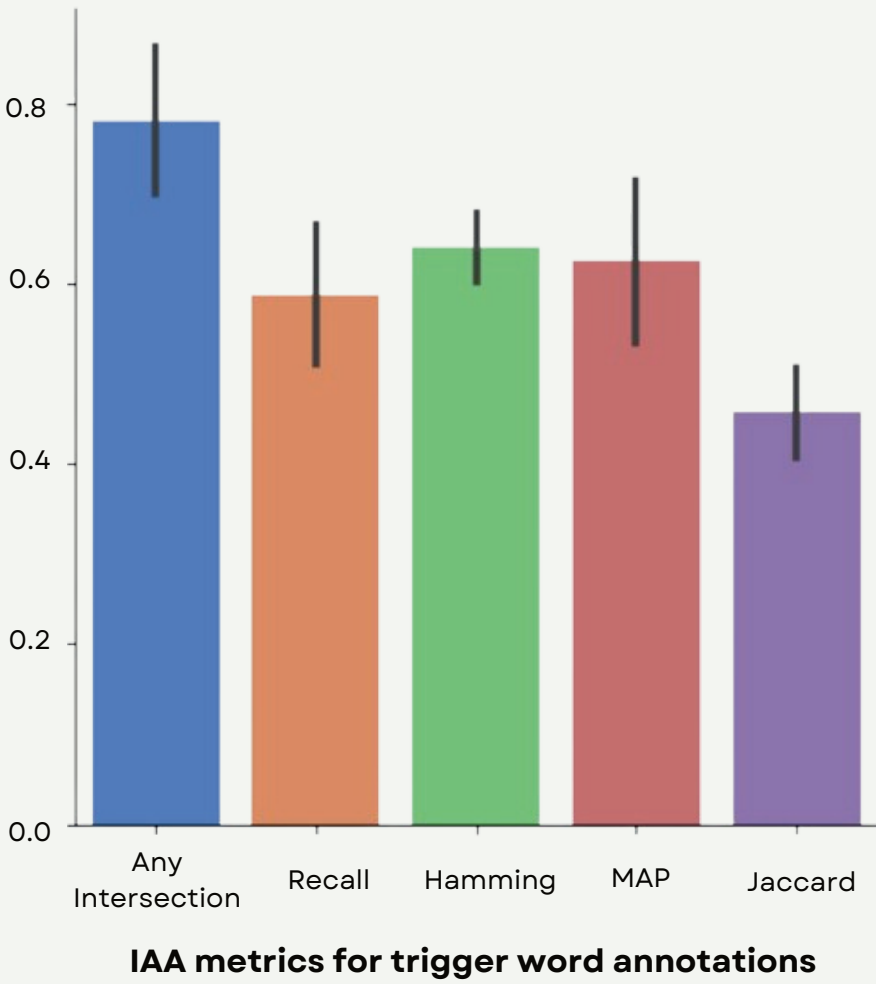
API = 95%

(NEW) Accumulated Precise Importance

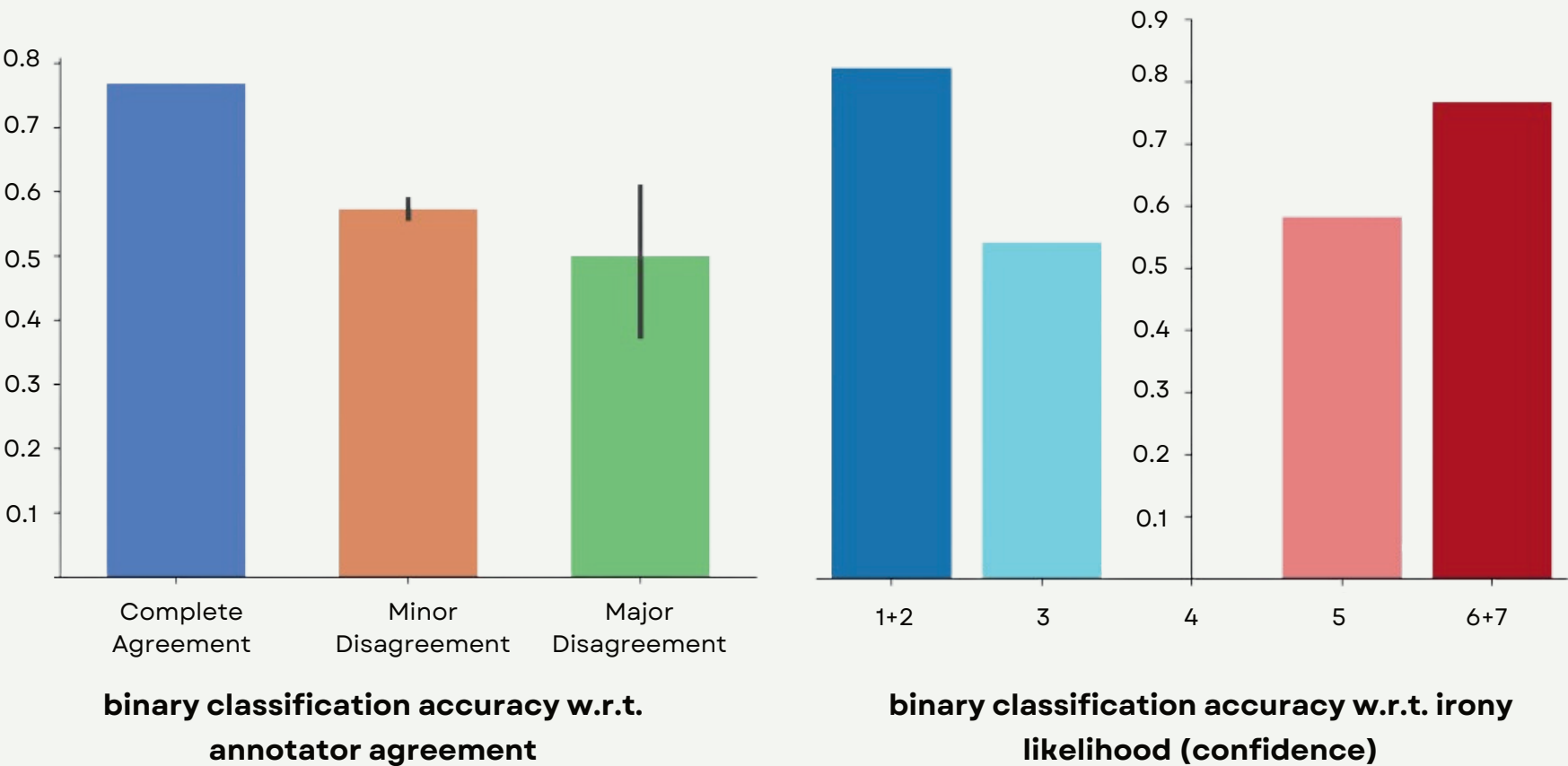
- for each human trigger word
- add up the system attribution
- result  $\approx$  sentence-level precision

## Corpus Stats

- 1 corpus:  
SemEval 2018 Task 3 (EN)
  - 2 irony likelihood agreement:  
Krippendorff's alpha: .87
  - 3 final label distribution:
- |   | 1     | 2  | 3   | 4  | 5   | 6   | 7     |
|---|-------|----|-----|----|-----|-----|-------|
| n | 2,778 | 26 | 179 | 77 | 271 | 437 | 1,024 |
- 4 trigger word agreement:  
complicated:



## Classification:



## Results:

**explainability:**  
40%+ attributions  
on irrelevant tokens

**classification:**  
better binary classification on  
high-confidence & high agreement

## Conclusion

- explainability for complex task has room for improvement.
- future work:  
train with fine-grained labels

