

Predicting Annotation Difficulty to Mitigate Annotator Bias

Study of Flemish Twitter Sentiment

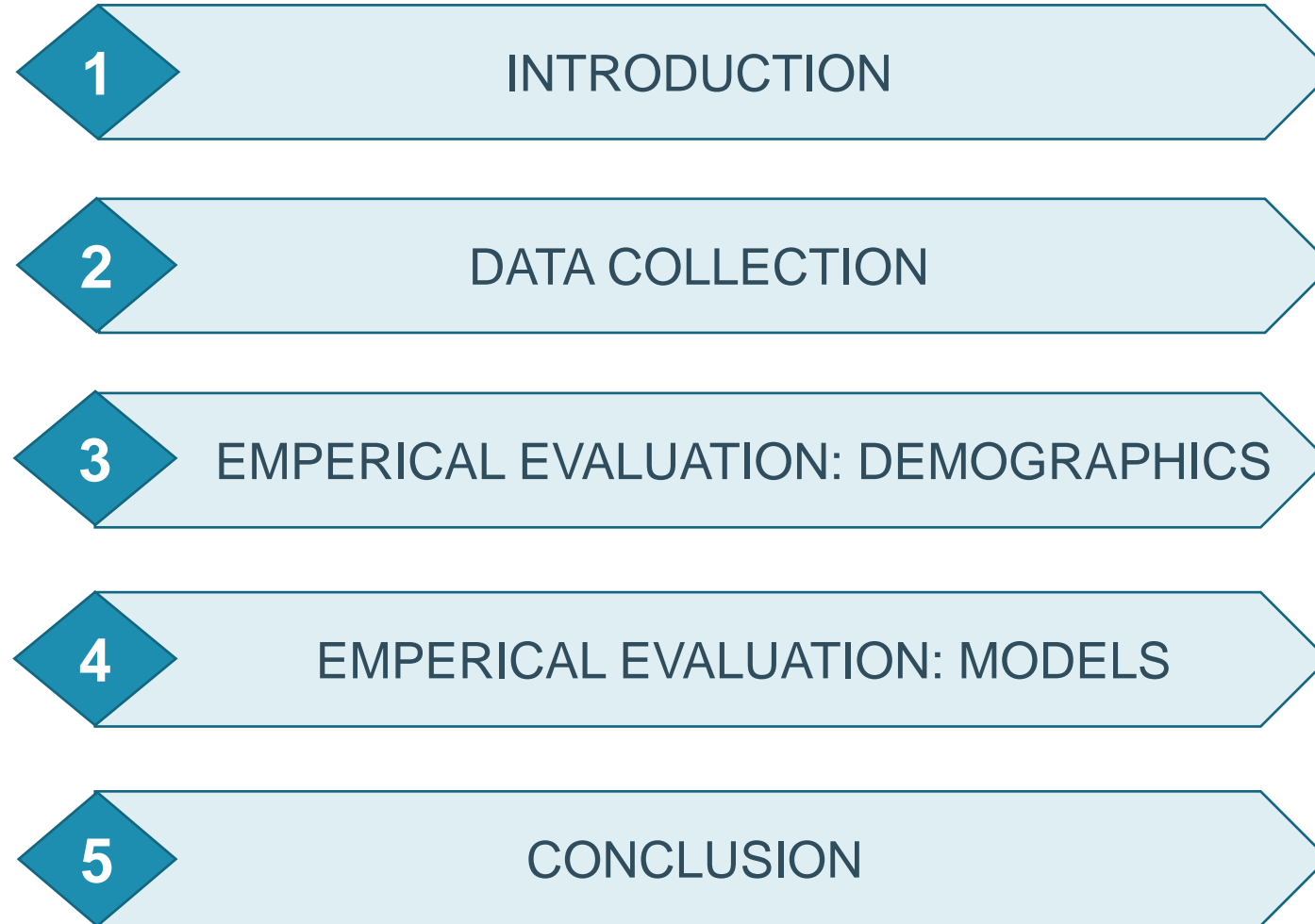
Joachim Bache-Mathiesen

Martijn Vanderschueren

Sophie De Waele



Overview



INTRODUCTION

1

INTRODUCTION

2

DATA COLLECTION

3

EMPIRICAL EVALUATION:
DEMOGRAPHICS

4

EMPIRICAL EVALUATION:
MODELS

5

CONCLUSION

Twitter Sentiment Analysis

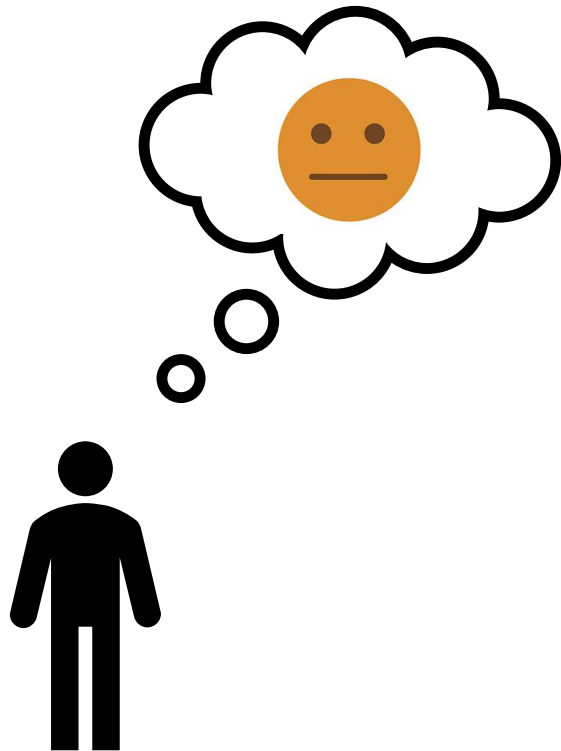


A screenshot of a Twitter tweet is shown, enclosed in a thick red border. The tweet is from user 'ianthe' (@Jeanthevh) and contains the text 'Te laat cola drinken is echt niet slim pfff...'. The tweet is dated '11:01 PM · Mar 8, 2016'. Below the text are icons for a heart (1), a speech bubble (Reply), and a link icon (Copy link). At the bottom of the tweet area is a blue button that says 'Explore what's happening on Twitter'. The Twitter logo is visible in the top right corner of the tweet area.

Twitter Sentiment Analysis



Twitter Sentiment Analysis



vtm nieuws VTM NIEUWS 
@VTMNIEUWS 

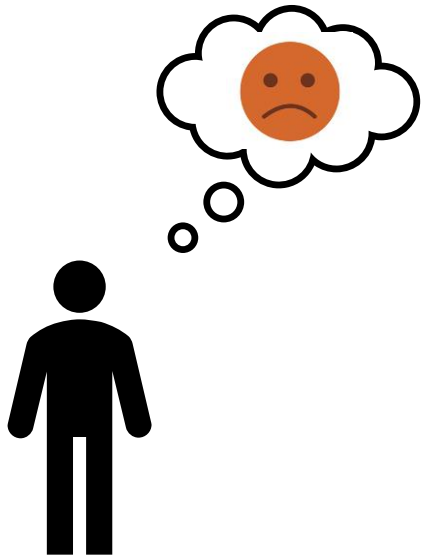
Om 8 uur wordt het overleg over CETA opnieuw hervat. Gisteravond bereikten de onderhandelaars geen akkoord. bit.ly/2eQzSVo

6:15 AM · Oct 26, 2016 

 1  Reply  Copy link

[Read 1 reply](#)

Twitter Sentiment Analysis: Problem I



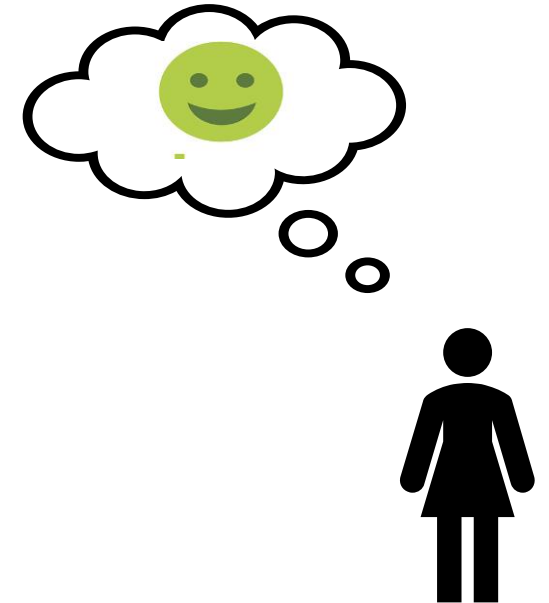
 **YØRBEN**
@YorbenDr 

Ik ben gta 5 op de pc aan het zoeken en vind het ni terug...
Life is great!

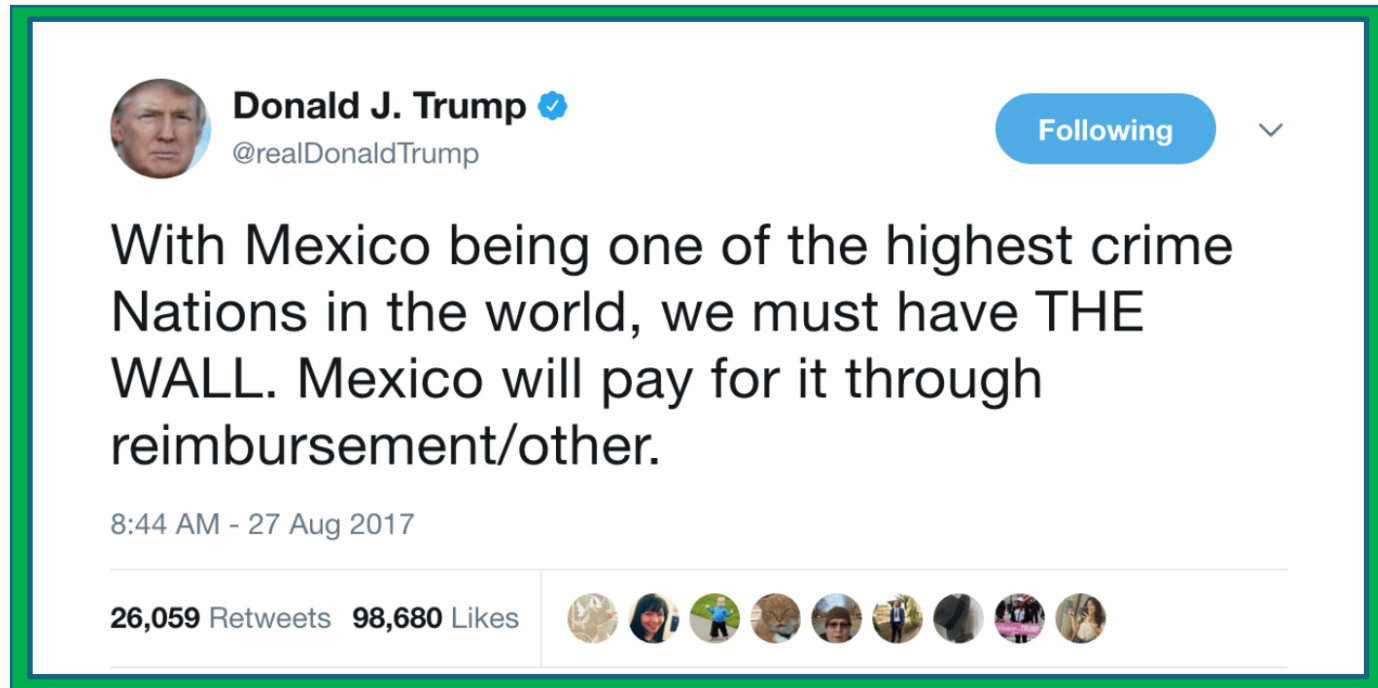
6:31 PM · May 9, 2016 

  Reply  Copy link

[Explore what's happening on Twitter](#)



Twitter Sentiment Analysis: Problem II



A screenshot of a tweet from Donald J. Trump (@realDonaldTrump) dated August 27, 2017. The tweet text is: "With Mexico being one of the highest crime Nations in the world, we must have THE WALL. Mexico will pay for it through reimbursement/other." The tweet has 26,059 retweets and 98,680 likes. The interface shows the user's profile picture, name, handle, and a "Following" button. The tweet is enclosed in a green border.



Annotator Bias



*A form of bias that annotators cause due to **different knowledge** in regards to the task and their **subjective perception***

Potential Solution: Crowdsourcing



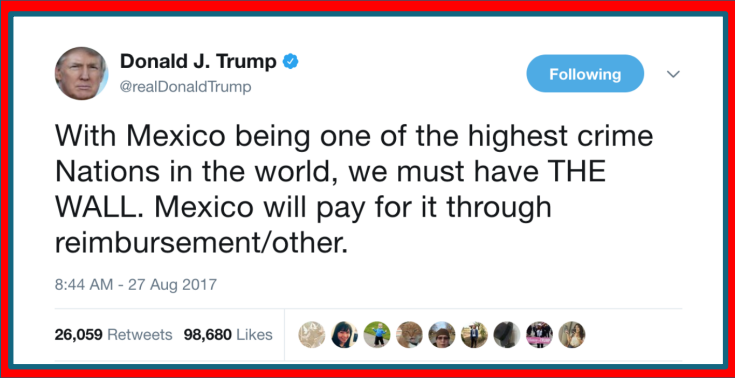
(Male, 33, Brugge, Highschool)



(Male, 19, Oostende, Highschool)



(Male, 39, Hasselt, Academic Bachelor)

A screenshot of a tweet by Donald J. Trump (@realDonaldTrump) dated 8:44 AM - 27 Aug 2017. The tweet text reads: "With Mexico being one of the highest crime Nations in the world, we must have THE WALL. Mexico will pay for it through reimbursement/other." The tweet has 26,059 Retweets and 98,680 Likes. The screenshot is framed with a red border.

(Male, 88, Asse, Highschool)

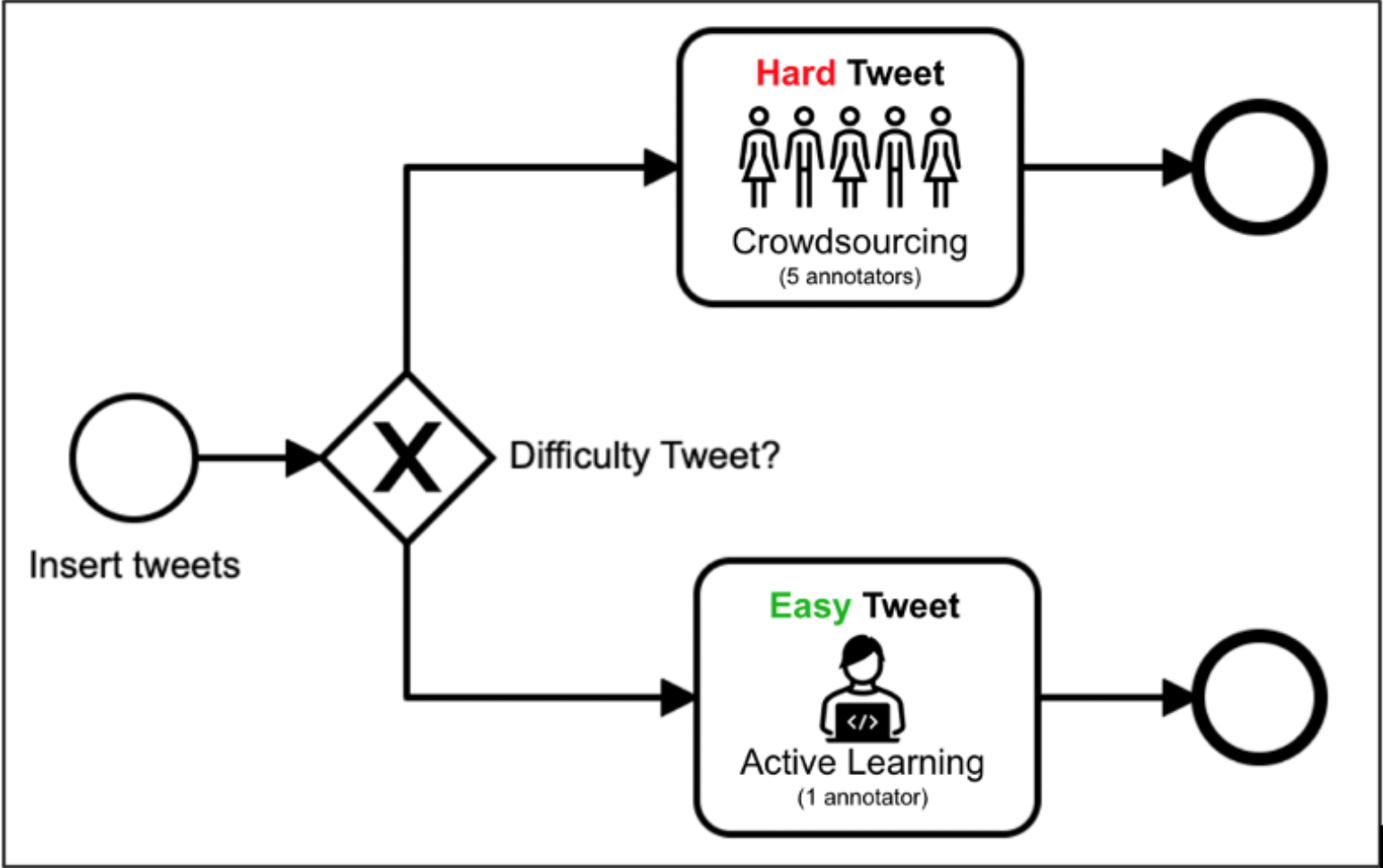


(Female, 22, Kortrijk, Academic Bachelor)



(Female, 80, Brussels, Professional Bachelor)

Proposed Scenario



Research Questions:

RQ: Which demographic factors **influence** annotator bias and how can it be **mitigated** efficiently?

-

SRQ1: Do annotators' **demographic features** such as gender, age and education **influence their annotations**?

-

SRQ2: Which machine learning models performs best in **mitigating** annotator bias by **identifying hard to annotate tweets**?

DATA COLLECTION

1

INTRODUCTION

2

DATA COLLECTION

3

EMPIRICAL EVALUATION:
DEMOGRAPHICS

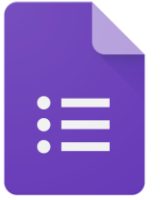
4

EMPIRICAL EVALUATION:
MODELS

5

CONCLUSION

Survey



Google Forms

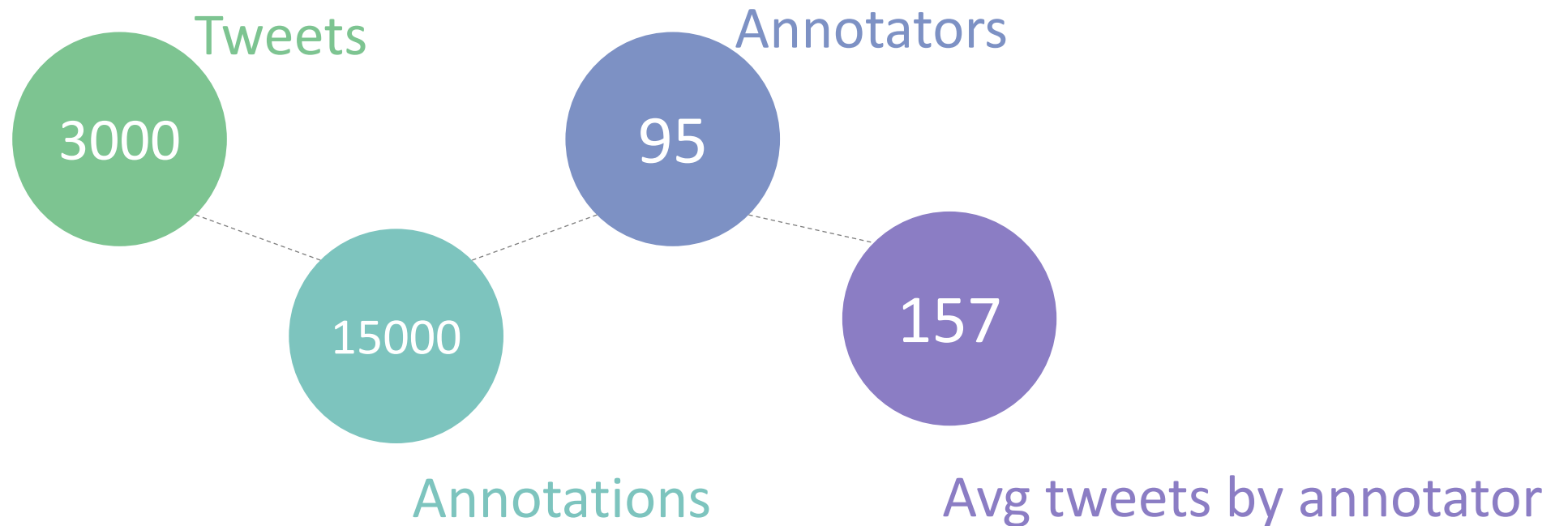
- 5 non-expert annotators for each tweet
- Annotators can fill in multiple different forms
- Sentiment of the author
- Labels:
 1. Positive
 2. Neutral
 3. Negative
 4. Unclear: don't understand the text

Als kind mag je toch dromen *

- positief
- neutraal
- negatief
- onduidelijk

Gathered data

- Age, Gender, Postcode and Education
- Forms were sent out to friends and family
- Manual with decision tree and examples



EMPERICAL EVALUATION: DEMOGRAPHICS

1

INTRODUCTION

2

DATA COLLECTION

3

EMPIRICAL EVALUATION:
DEMOGRAPHICS

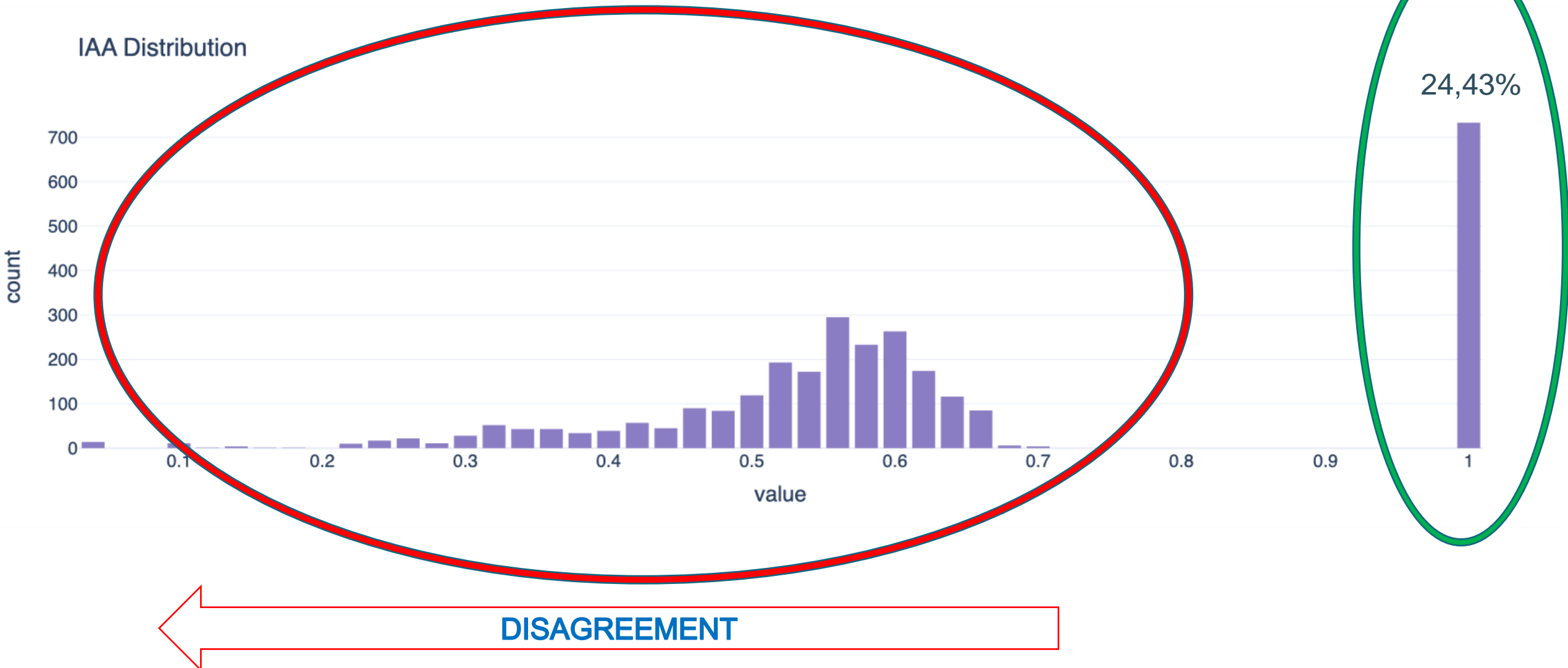
4

EMPIRICAL EVALUATION:
MODELS

5

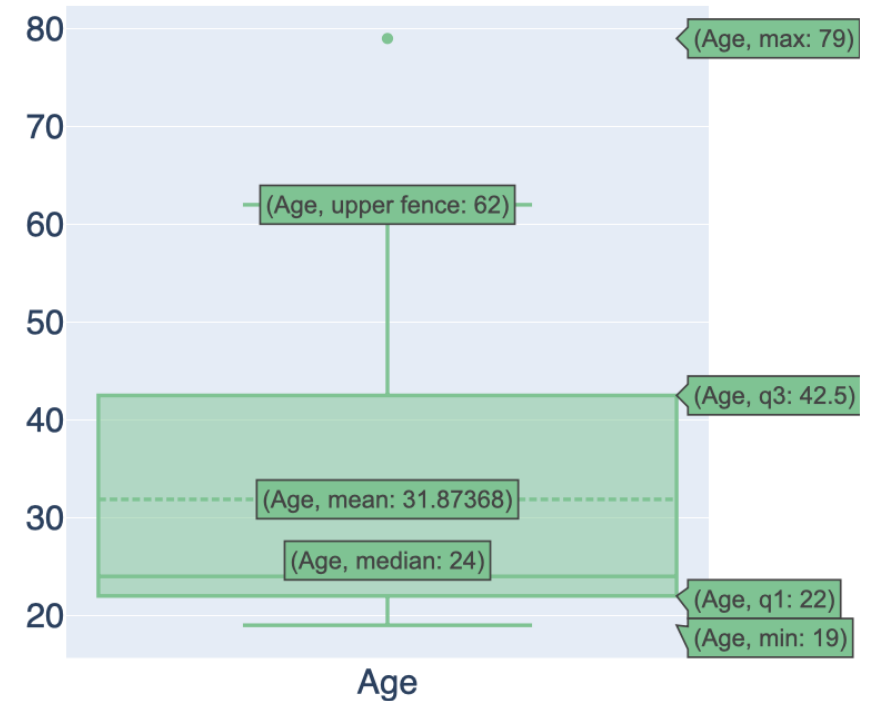
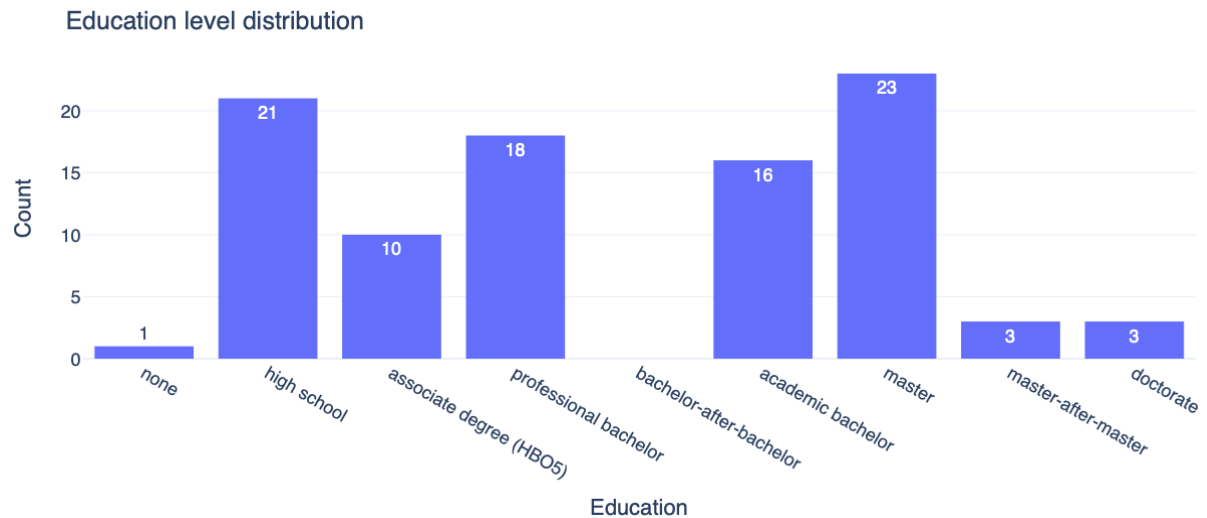
CONCLUSION

Inter-annotator agreement (IAA) distribution



Collected data

- Gender: \pm equal
- Age: overrepresentation of young adults
- Education: no university and university
- Province: Mainly West-Vlaanderen and Vlaams-Brabant



Agreement demographic groups

Cohen's Kappa

- Compare majority label within one tweet
- Gender, age and education
- Moderate agreement

Proportion test

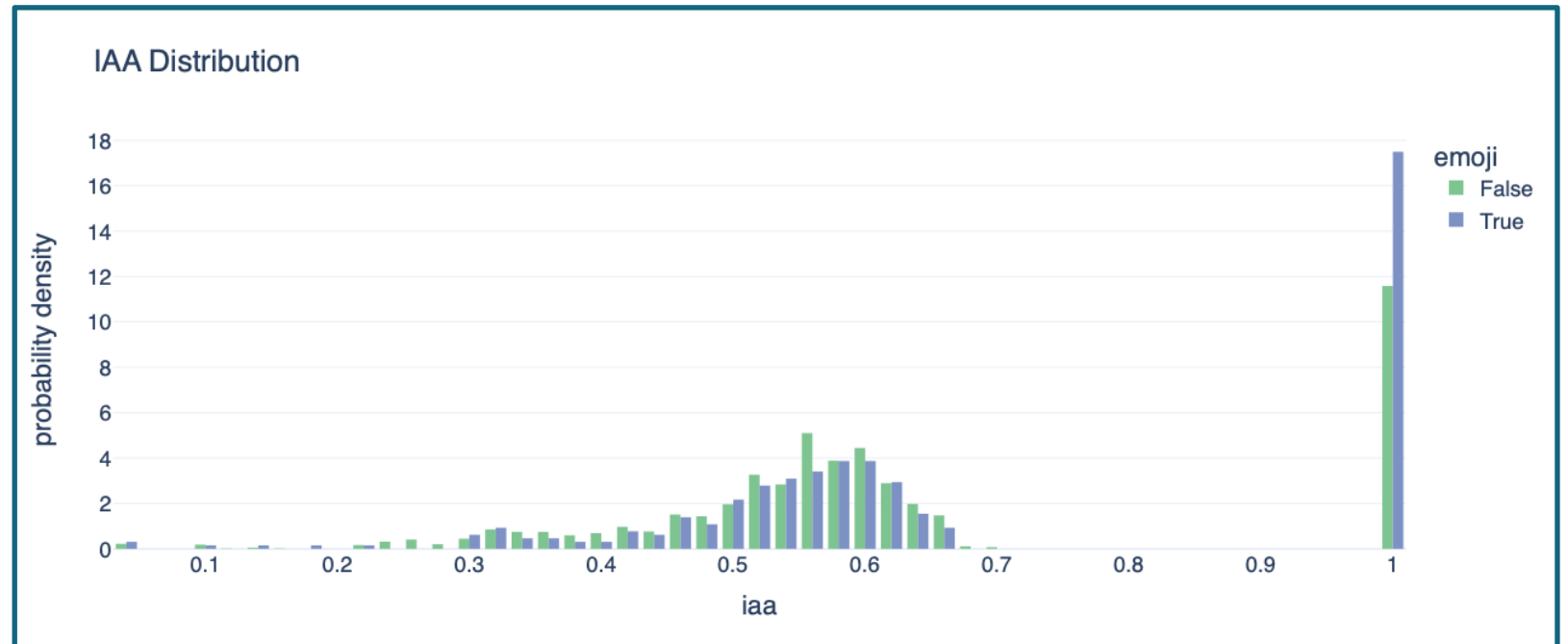
- Compare score within one group
- Gender, age and education
- Significant difference



Demographics should be considered when hiring annotators

Emojis

- More agreement when tweet contains emoji?
- Mann-Whitney U test
- Significant
- Input feature



EMPERICAL EVALUATION: MODELS

1

INTRODUCTION

2

DATA COLLECTION

3

EMPIRICAL EVALUATION:
DEMOGRAPHICS

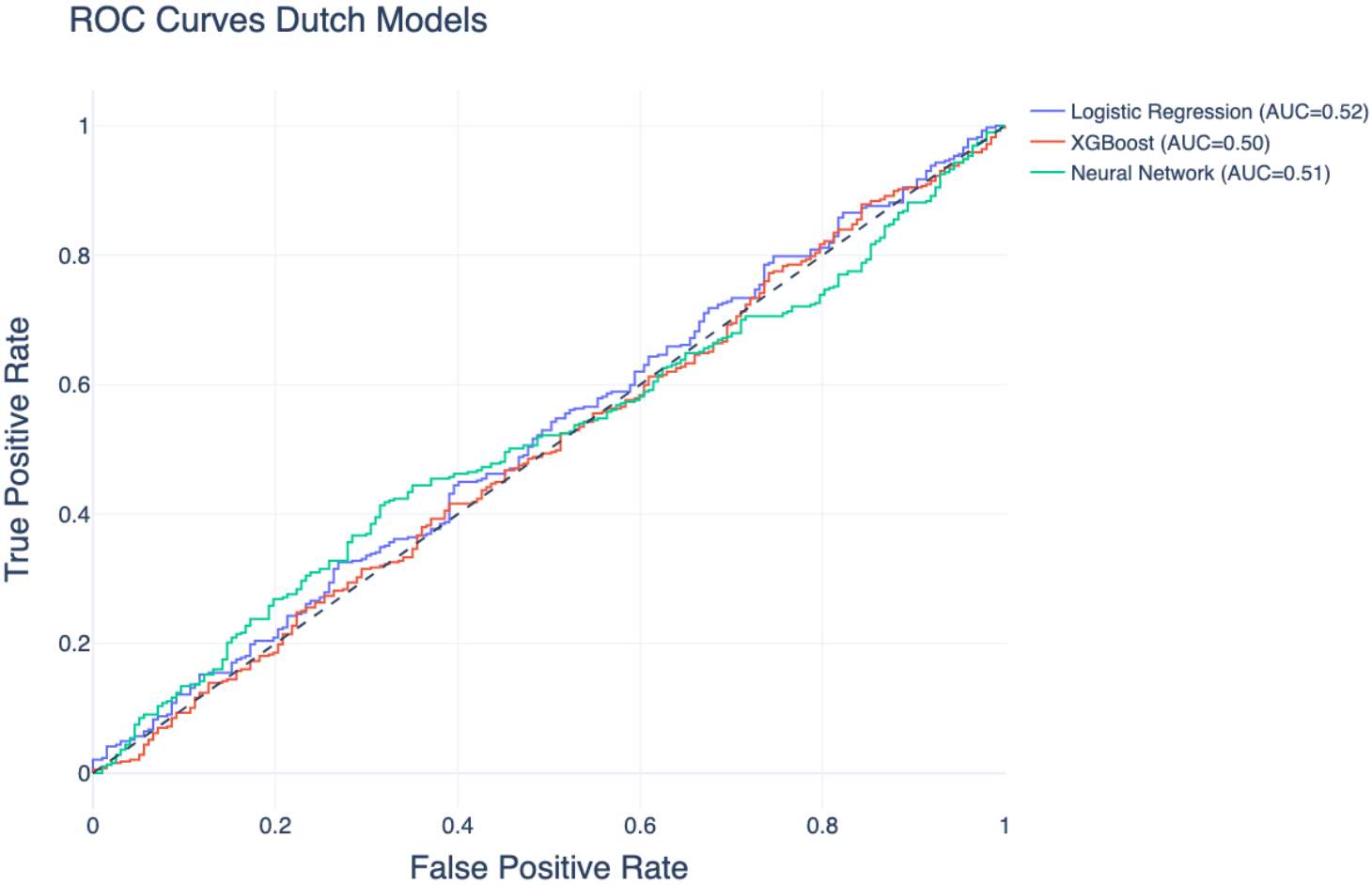
4

EMPIRICAL EVALUATION:
MODELS

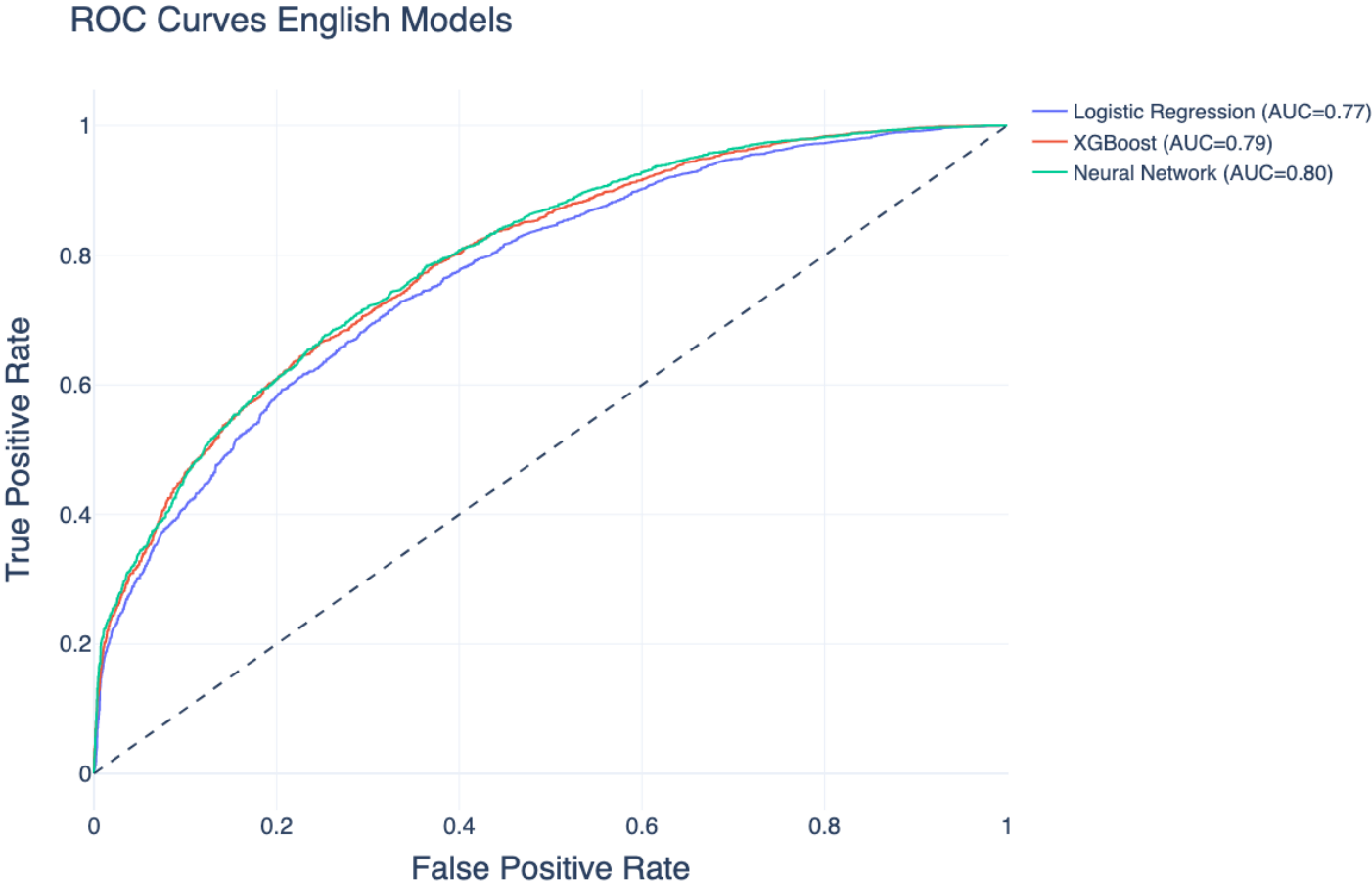
5

CONCLUSION

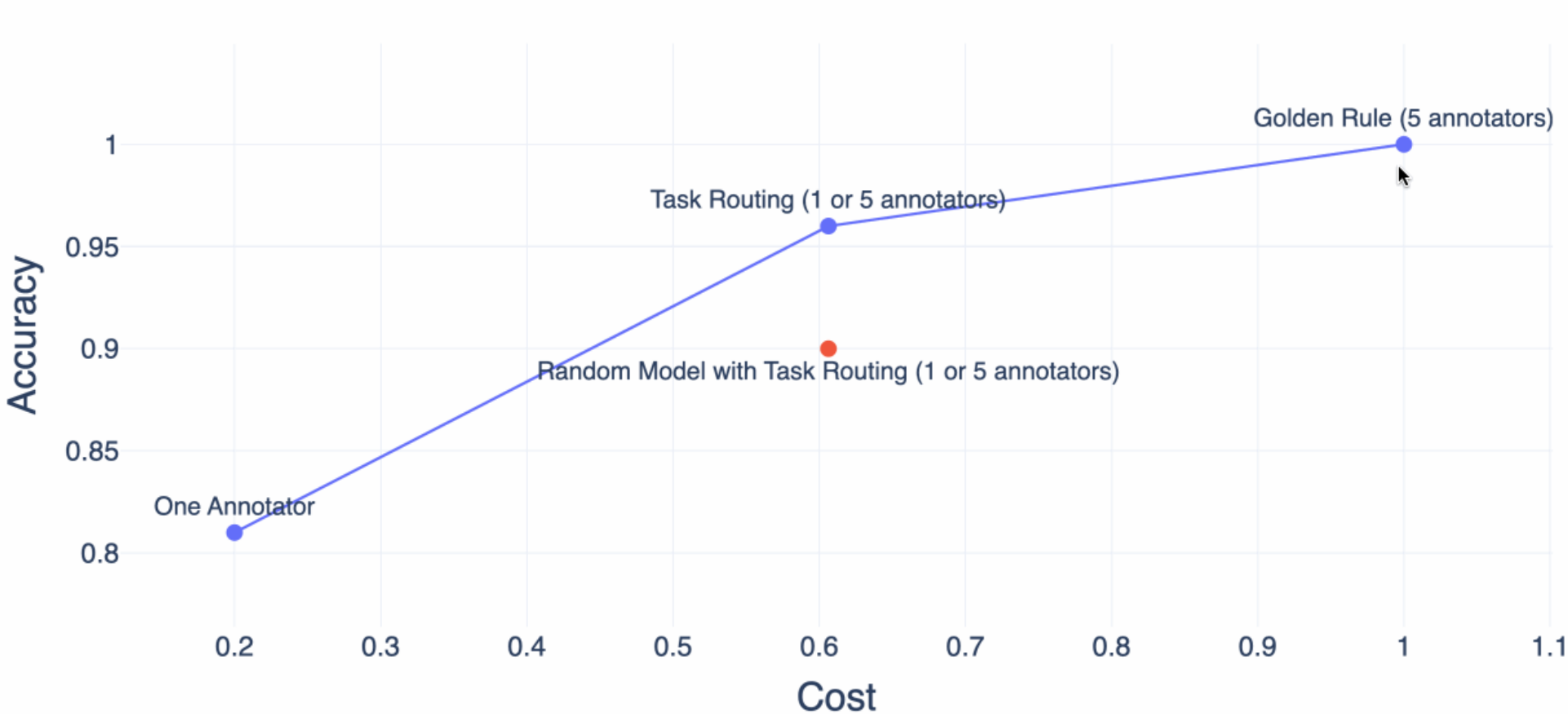
Model Evaluation



Model Evaluation



Model Evaluation



CONCLUSION

1

INTRODUCTION

2

DATA COLLECTION

3

EMPIRICAL EVALUATION:
DEMOGRAPHICS

4

EMPIRICAL EVALUATION:
MODELS

5

CONCLUSION

Conclusion

Sub-research question I:

Do annotators' demographic features such as gender, age and education influence their annotations?

- Consider demographic features in hiring annotators

Sub-research question II:

Which machine learning models performs best in mitigating annotator bias by identifying hard to annotate tweets?

- Neural network
- Relatively performant in selecting hard to annotate tweets
- Cost-efficient gathered data set while giving up relatively little in accuracy



Consequently, annotator bias can be mitigated efficiently

QUESTIONS?