Machine Translation quality across demographic dialectical variation in Social Media

Adi Renduchintala and Dmitriy Genzel
Facebook AI

Biases in Machine Learning

- Machine learning systems can encode harmful societal biases.
- Widespread use of machine learning systems amplify these biases.

Biases in Machine Learning (in NLP)

- Machine learning systems can encode harmful societal biases
- Widespread use of machine learning systems amplify these biases.

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$$

Bolukbasi et al. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, Advances in neural information processing systems, 2016

Biases in Machine Learning (in Vision)

- Machine learning systems can encode harmful societal biases
- Widespread use of machine learning systems amplify these biases.

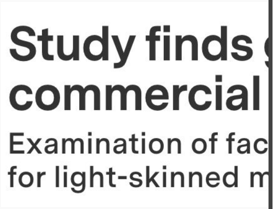
Study finds gender and skin-type bias in commercial artificial-intelligence systems

Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.

http://gendershades.org/ &
news.mit.edu

Biases in Machine Learning (in Vision)

- Machine learning systems can encode harmful societal biases
- Widespread use of machine learning systems amplify these biases.



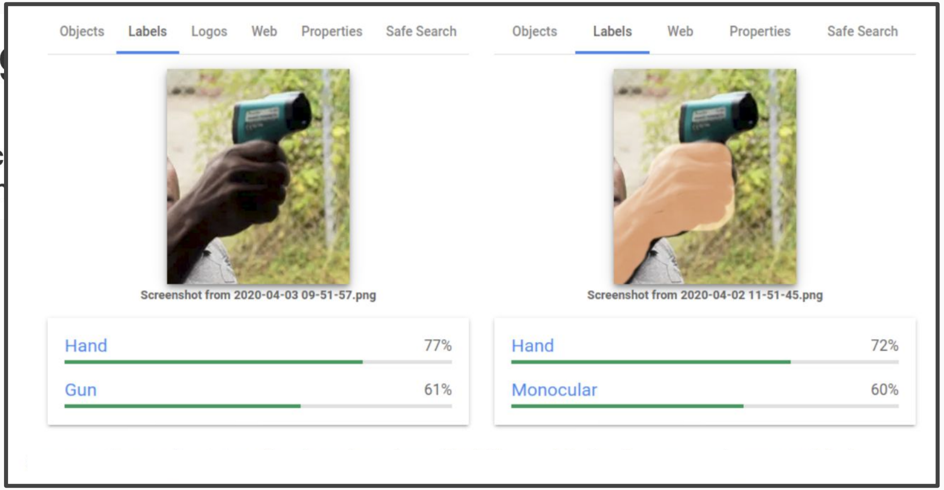


Image Credit:
@bjnagel & algorithmwatch.org

Biases in Machine Learning (in Vision)

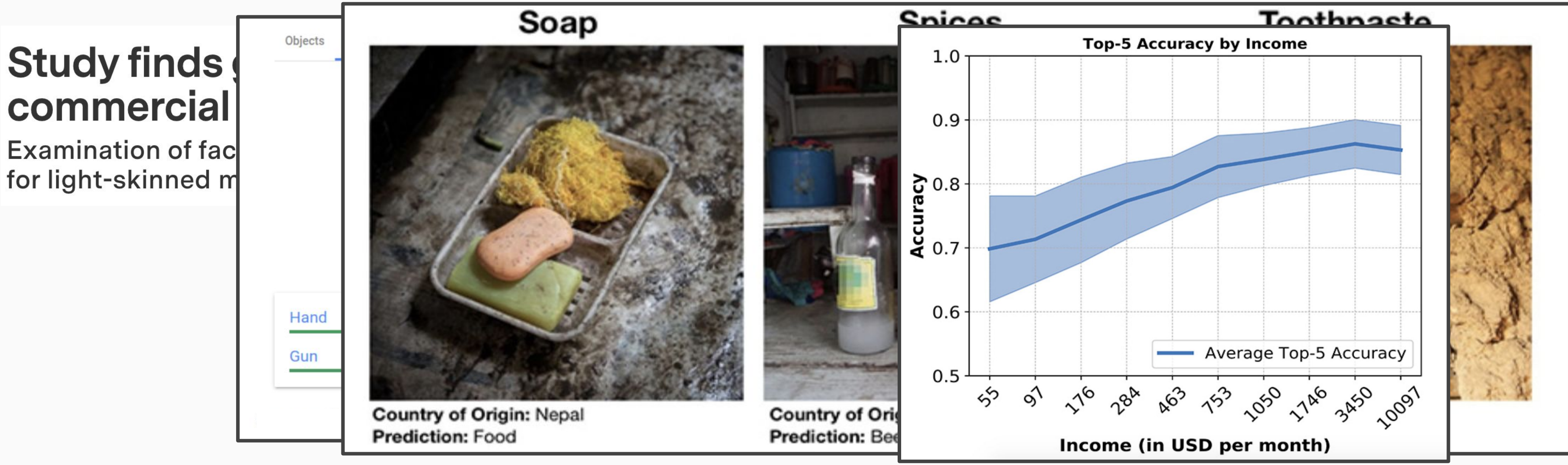
- Machine learning systems can encode harmful societal biases
- Widespread use of machine learning systems amplify these biases.

Biases in Machine Learning (in Vision)

- Machine learning systems can encode harmful societal biases
- Widespread use of machine learning systems amplify these biases.

Biases in Machine Learning (in Vision)

- Machine learning systems can encode harmful societal biases
- Widespread use of machine learning systems amplify these biases.



Biases in Machine Learning (in ASR)

- Machine learning systems can encode harmful societal biases
- Widespread use of machine learning systems amplify these biases.

There Is a Racial Divide in Speech-Recognition Systems, Researchers Say

Technology from Amazon, Apple, Google, IBM and Microsoft misidentified 35 percent of words from people who were black. White people fared much better.

Biases in Machine Learning (in MT?)

- Machine learning systems can encode harmful societal biases
- Widespread use of machine learning systems amplify these biases.

Goal: Investigate if modern machine translation systems amplify racial biases?

Proposal

- Use twitter posts which have demographic dialect information associated.
- Translate these tweets with 3 “off-the-shelf” machine translation models
- Do we notice disparity in translation quality?

Data

- We use data that was released in **prior work** by:
  - Blodgett, et al. *Demographic dialectal variation in social media: A case study of African-American English.* EMNLP, 2016
- This data was automatically annotated with racial dialectal labels by the same authors.

Data

- We use data that was released in prior work by:
  - Blodgett, et al. *Demographic dialectal variation in social media: A case study of African-American English.* EMNLP, 2016
- This data was ⚠️ **automatically annotated** ⚠️ with racial dialectal labels by the same authors.

Data

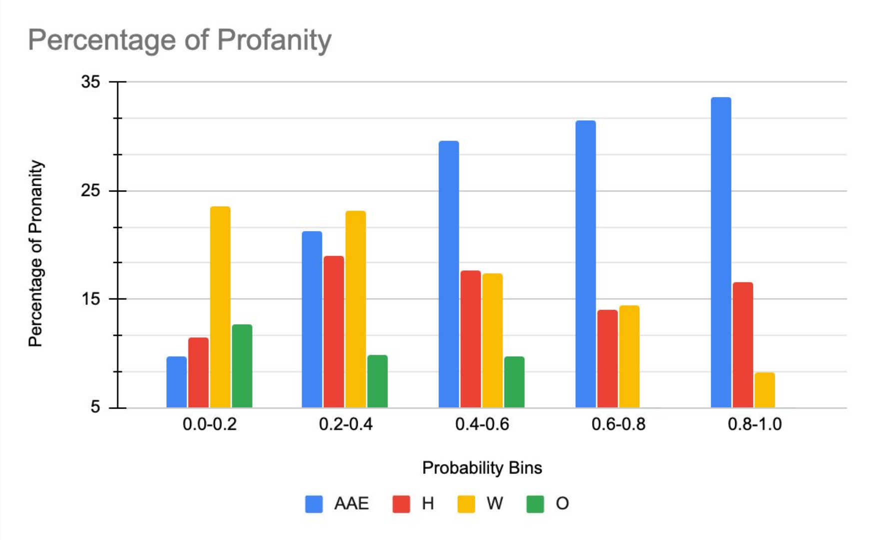
- We use data that was released in prior work by:
  - Blodgett, et al. *Demographic dialectal variation in social media: A case study of African-American English.* EMNLP 2016
- This data was ⚠️ **automatically annotated** ⚠️ with racial dialectal labels by the same authors.
  - A weakly supervised mixed-membership model was used.
  - The authors generated a posterior distribution over 4 categories for each tweet:
    - African-American English (AAE)
    - Hispanic English (H)
    - White-aligned English (W)
    - Other

Data

Examples	**AAE**	**H**	**W**
Either yu gone get yo fkn life or get out my fkn life	0.82	0.004	0.142
When you got somebody good, you hold on to ' em .	0.45	0.016	0.527
My sister asked me if the lions are in the playoffs..	0.011	0.023	0.965
I'm too sad to stay up and im tired and i have church so night	0.006	0.873	0.12

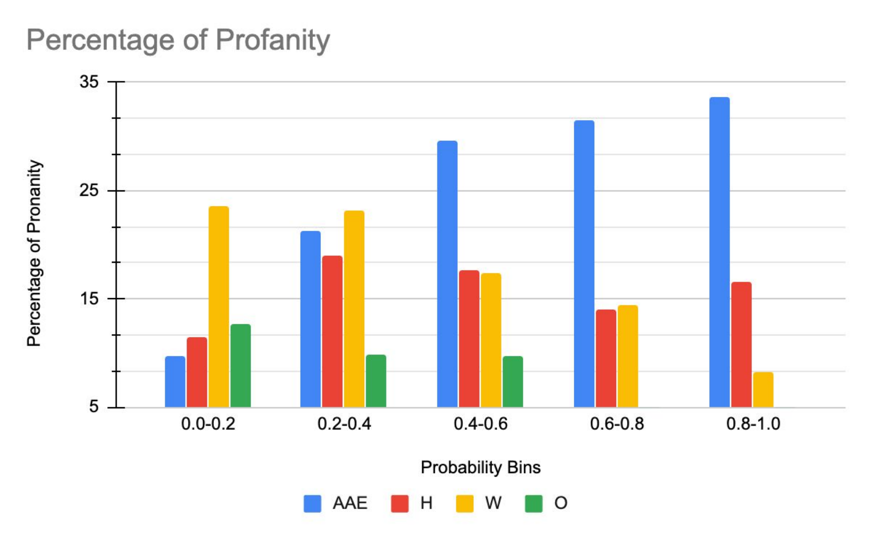
Profanity and Predictions

- The weakly supervised model seems to think that profanity is a feature of the AAE dialect.
- This is not observed in any of the other dialects.
- we filter out all tweets with profanity, to not be influenced by the weakly supervised model's (potentially) spurious correlations.



Data Challenges

- The dataset definitely has some flaws (correlating profanity with a demographic dialect is one example)
- However, the lack of expert annotated data to conduct analysis of this nature is also an issue.

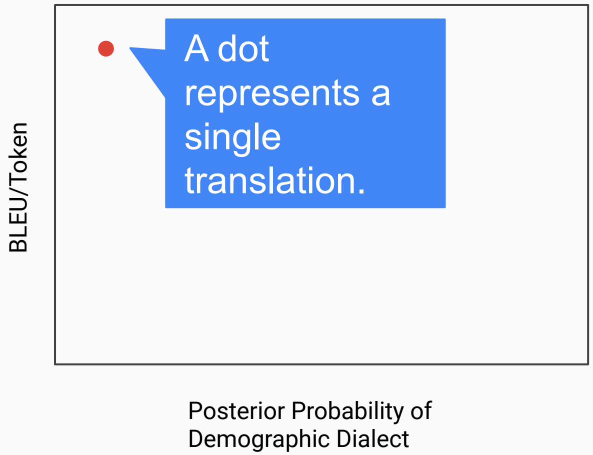


Experimental Setup

- For each category we subdivide the tweets into 5-bins based on the posterior probability (0.0 - 0.2, 0.2 - 0.4, … 0.8 - 1.0)
- From each bin in each category we sample ~30 tweets and have then translated into French by professional translators.
- We then used 3 “off-the-shelf” translation systems to translate the ~600 tweets using an English->French model.
- We plot the quality of the translation against the posterior probability of being a demographic category.

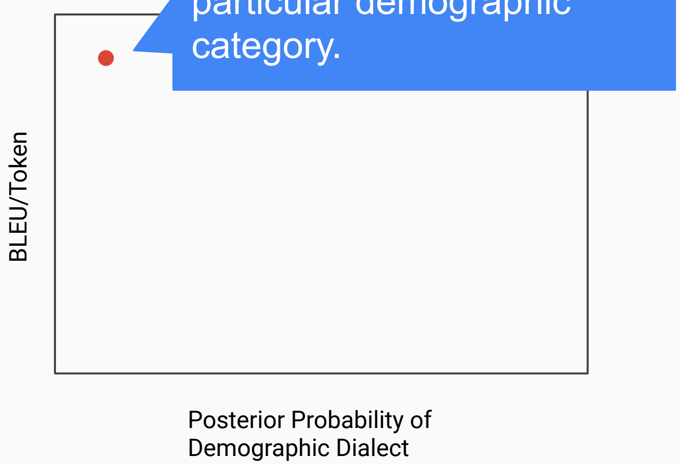
Results

- We plot BLEU/ (num. Reference-tokens) along the y-axis and the posterior probability of the tweet belonging to a demographic dialect category.



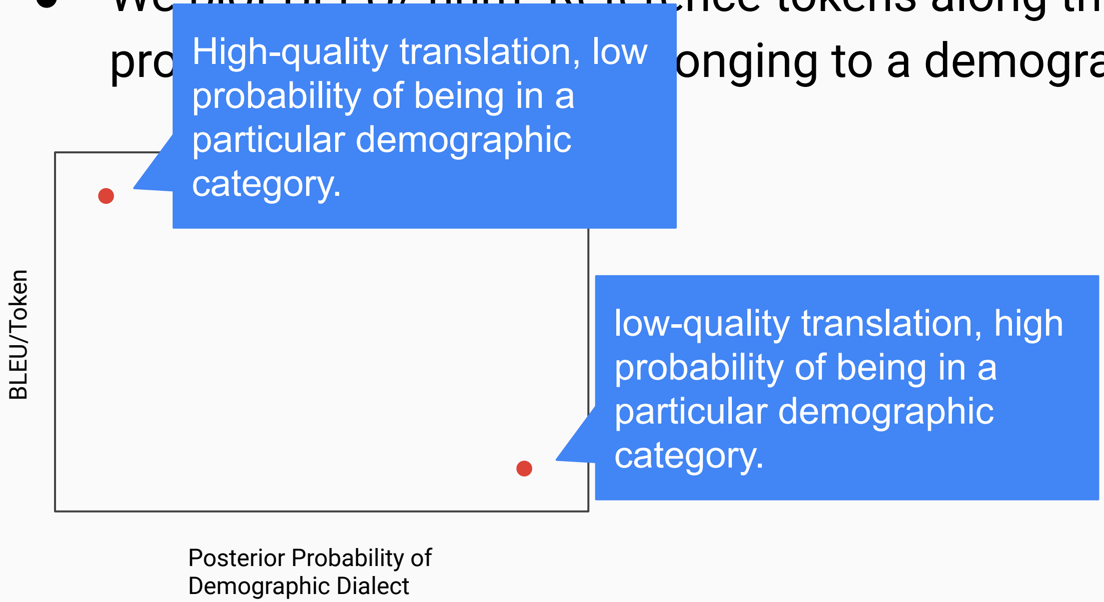
Results

- We plot BLEU/ num. Reference-tokens along the y-axis and the posterior pro onging to a demographic dialect category.



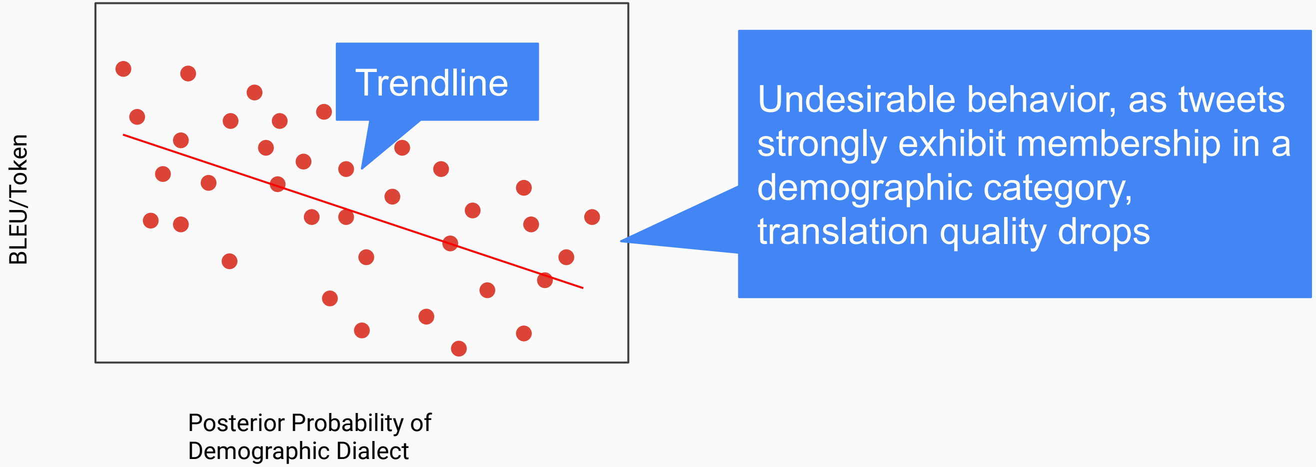
Results

- We plot BLEU/ num Reference-tokens along the y-axis and the posterior pro onging to a demographic dialect category.



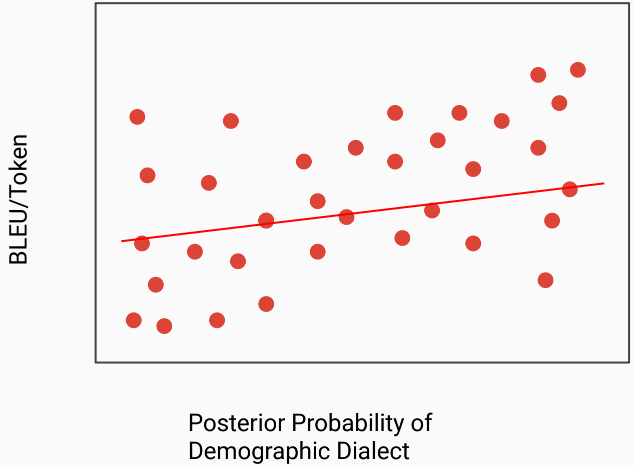
Results

- We plot BLEU/ num. Reference-tokens along the y-axis and the posterior probability of the tweet belonging to a demographic dialect category.



Results

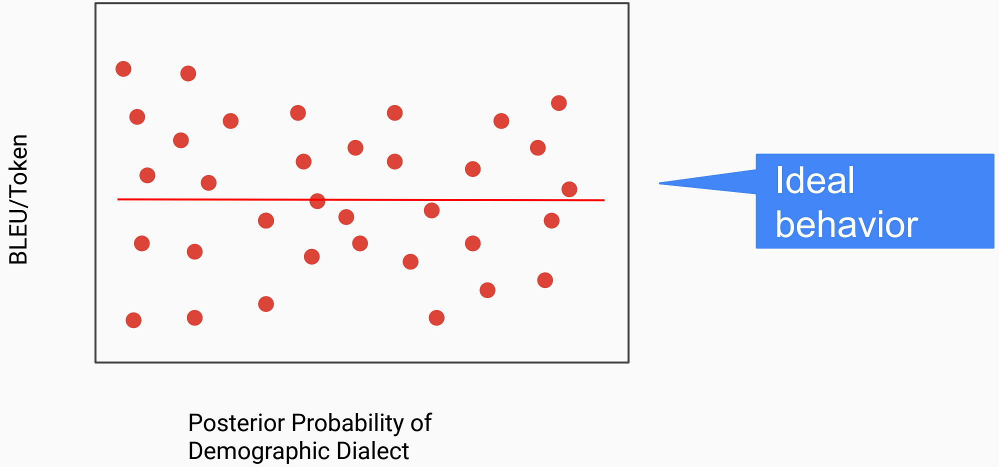
- We plot BLEU/ num. Reference-tokens along the y-axis and the posterior probability of the tweet belonging to a demographic dialect category.



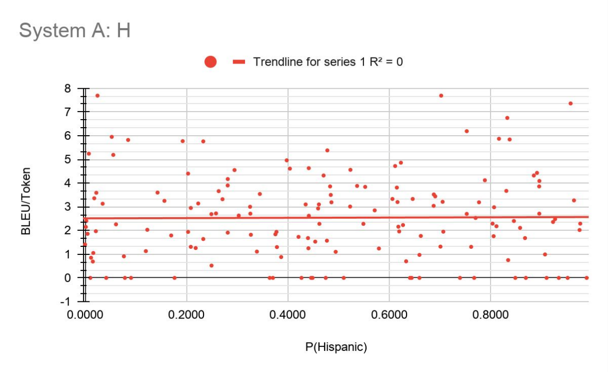
Also undesirable behavior

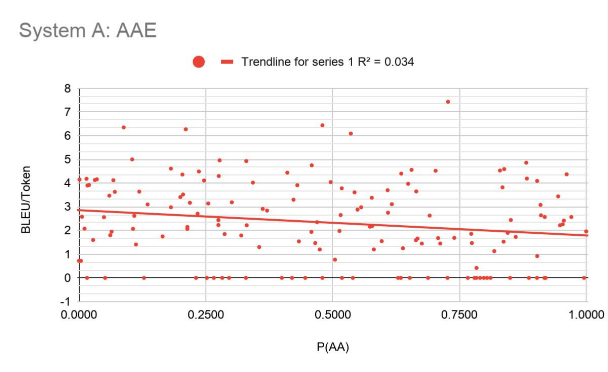
Results

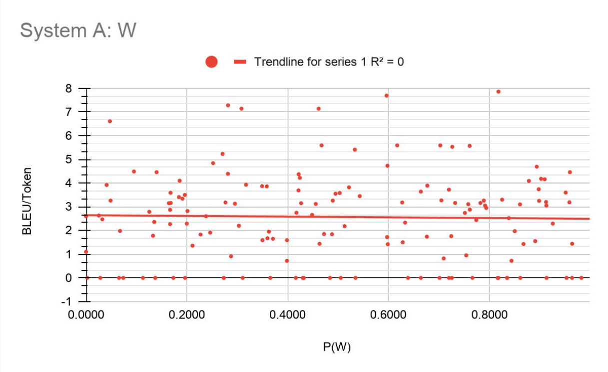
- We plot BLEU/ num. Reference-tokens along the y-axis and the posterior probability of the tweet belonging to a demographic dialect category.



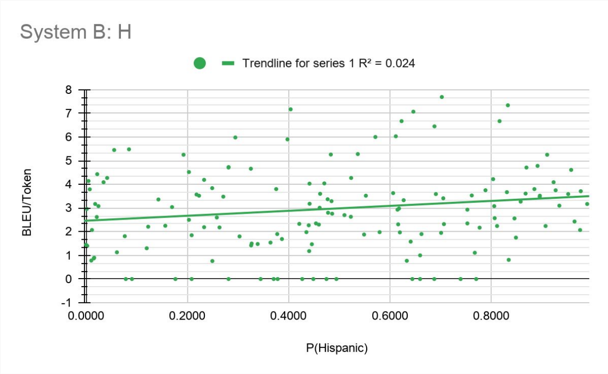
Results

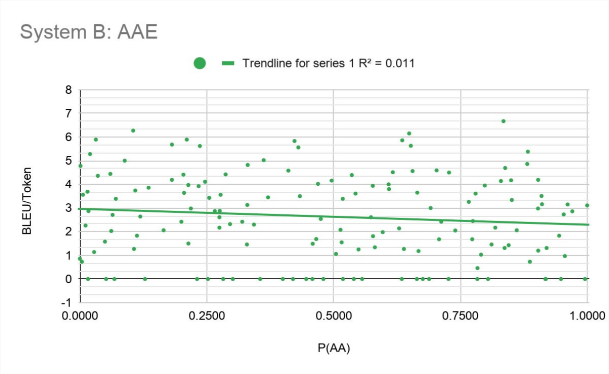

System A: H
Trendline for series 1 R² = 0
BLEU/Token
P(Hispanic)

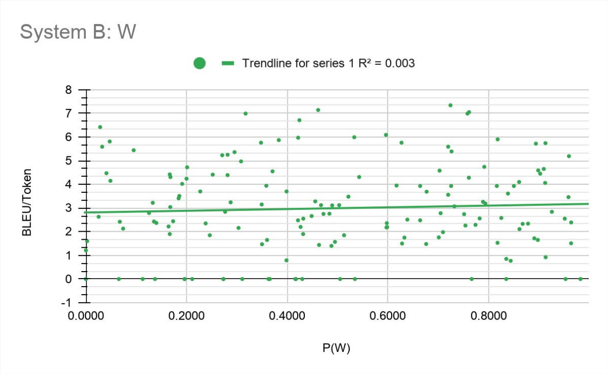

System A: AAE
Trendline for series 1 R² = 0.034
BLEU/Token
P(AA)


System A: W
Trendline for series 1 R² = 0
BLEU/Token
P(W)

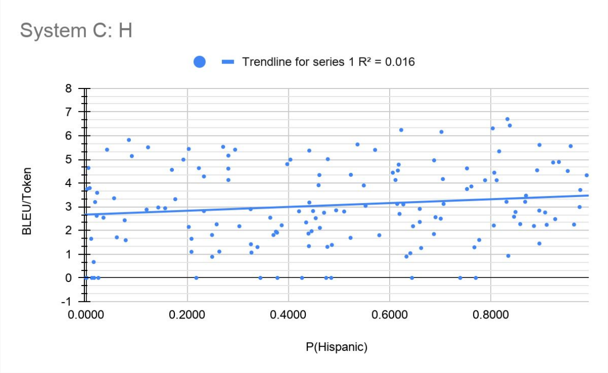
Results

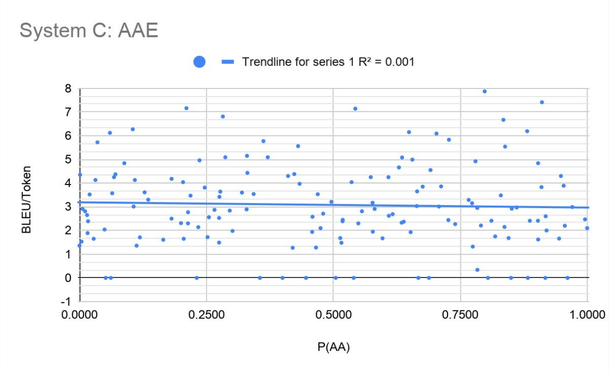

System B: H
Trendline for series 1 R² = 0.024
BLEU/Token
P(Hispanic)
8
7
6
5
4
3
2
1
0
-1
0.0000
0.2000
0.4000
0.6000
0.8000

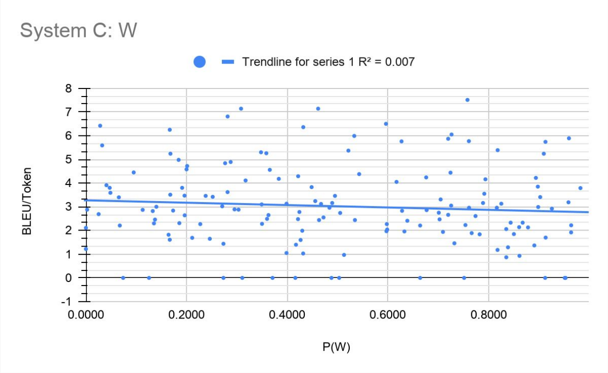

System B: AAE
Trendline for series 1 R² = 0.011
BLEU/Token
P(AA)
8
7
6
5
4
3
2
1
0
-1
0.0000
0.2500
0.5000
0.7500
1.0000


System B: W
Trendline for series 1 R² = 0.003
BLEU/Token
P(W)
8
7
6
5
4
3
2
1
0
-1
0.0000
0.2000
0.4000
0.6000
0.8000

Results


System C: H
Trendline for series 1 R² = 0.016
BLEU/Token
P(Hispanic)
8
7
6
5
4
3
2
1
0
-1
0.0000
0.2000
0.4000
0.6000
0.8000


System C: AAE
Trendline for series 1 R² = 0.001
BLEU/Token
P(AA)
8
7
6
5
4
3
2
1
0
-1
0.0000
0.2500
0.5000
0.7500
1.0000


System C: W
Trendline for series 1 R² = 0.007
BLEU/Token
P(W)
8
7
6
5
4
3
2
1
0
-1
0.0000
0.2000
0.4000
0.6000
0.8000

Conclusion

- Our experiments suggest that modern NMT systems exhibit undesirable behavior when dealing with input associated with AAE dialects.
- Further work is needed to understand this phenomenon better. Ideally, analysis should be conducted on expert annotated data.
- Our hope is that this work is a call to action to consider this a serious problem and mitigate the amplification of biases via AI systems.
- One concrete recommendation is to include analysis like this into model evaluation.