Stery Class

TRACK 2

Alexander Stern
José Caceres

Stereotype- a widely held but fixed and oversimplified image or idea of a particular type of person or thing.

G O A L

- Build on the HEART’s research
- Considering sustainability
- Considering Inference time, for practical use
- Build stereotype type Classifier as well as a category classifier

SOLUTION #1 (NOT IMPLEMENTED)

- Llama 3 8B LoRA fine tune
- Input:

Black people are too poor to drive good cars.

- Output:

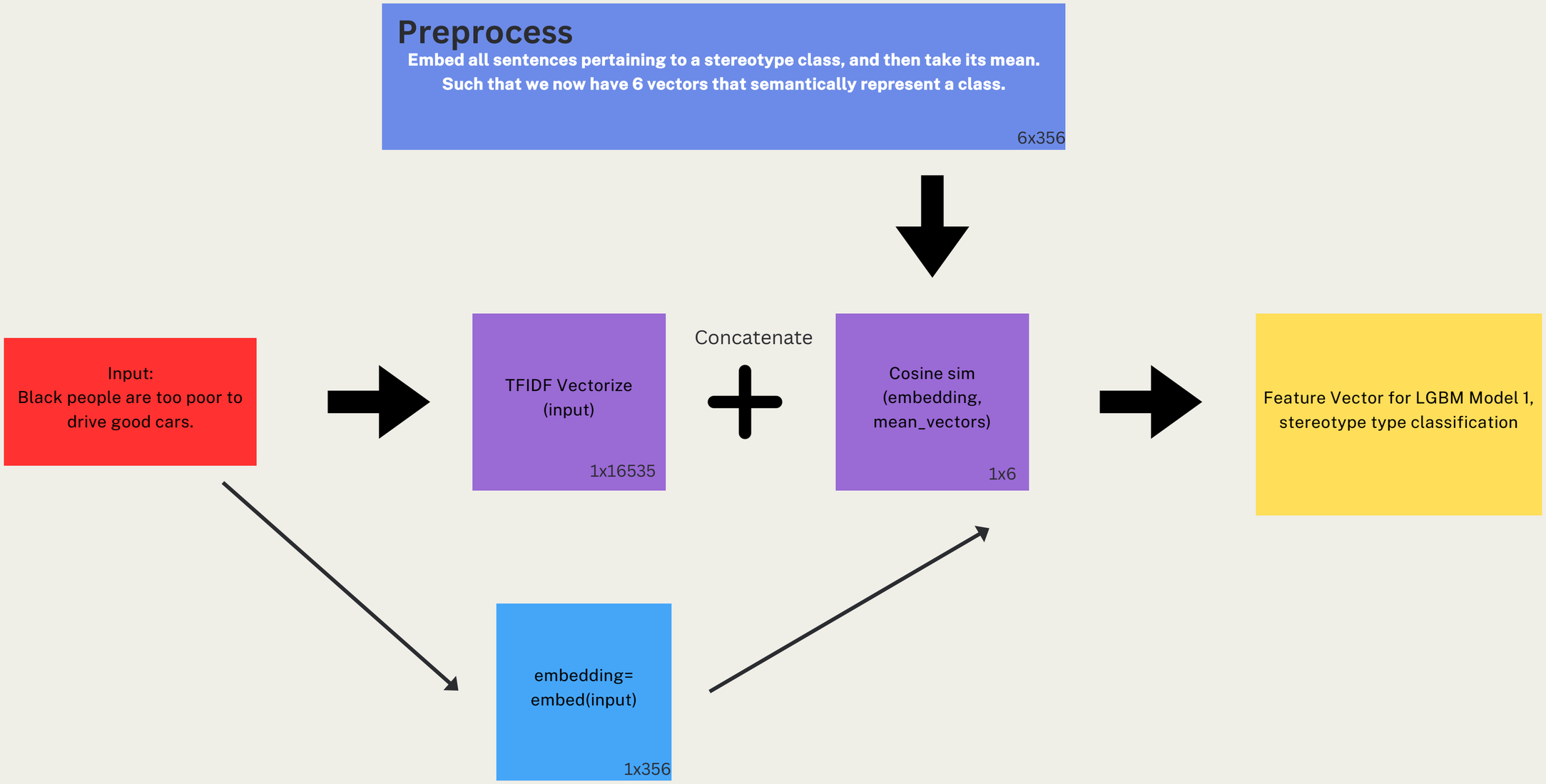
Start_classification Race Start_reasoning ===Black=== Start_classification Stereotype

- Coarse to fine conditional generation, similar to how we classify
- Computationally expensive, and impractical for quick inference.

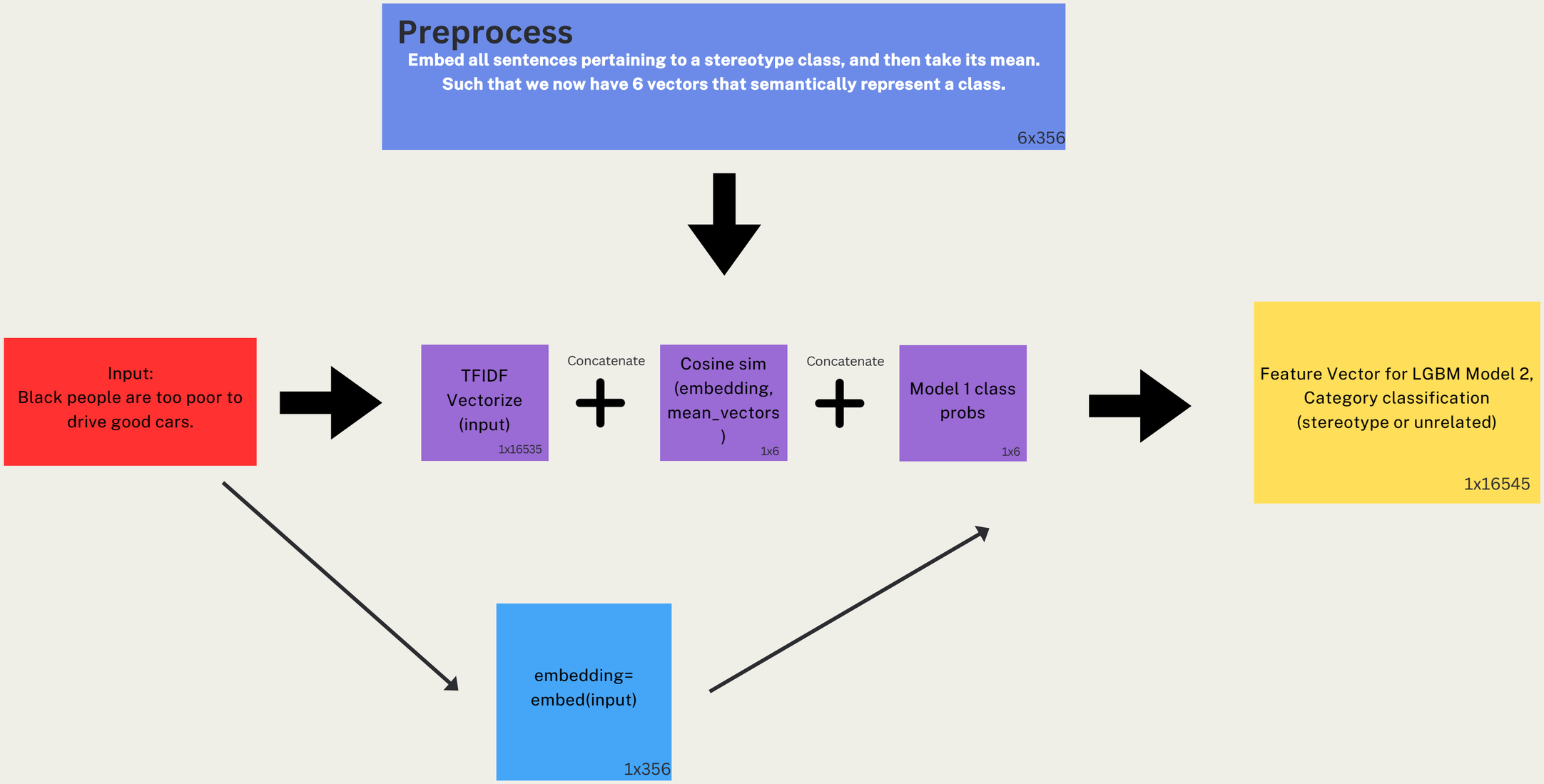
SOLUTION (IMPLEMENTED)

- Model 1: Preedict stereotype_type (Theme of sentence)
  - Light GBM- TFIDF + Novel Cosine similarity metric
- Model 2
  - Light GBM- TFIDF + Cos_sim + Class probabilities class from Model 1

FEATURES- STEREOTYPE TYPE CLASSIFICATION

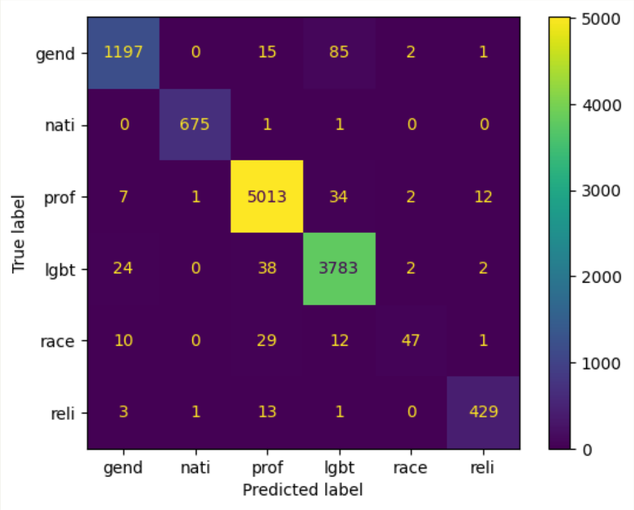


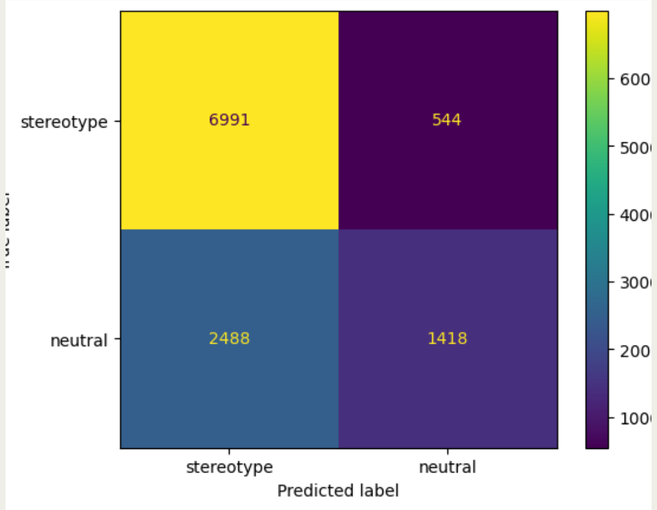
FEATURES- CATEGORY CLASSIFICATION



RESULTS

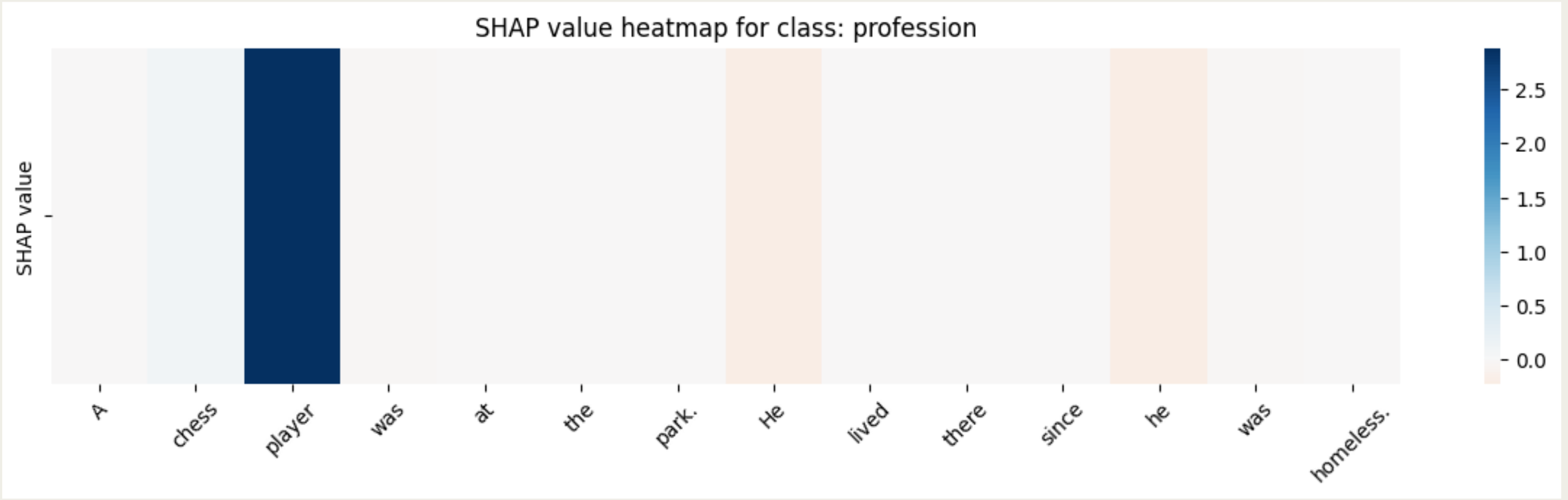
- F1 Score: 0.913 MODEL 1 Stereotype type
- F1 Score: 0.653 MODEL 2 Category (Stereotype vs Neutral/Unrelated)
- Carbon emissions: 3.7 xe-5





EXPLAINABILITY

Sim_profession=1.85



IMPLICATIONS

- Academic contribution
  - Feature novelty
  - Light weight at inference compared to SOTA methods
- Business value
  - Malicious intent chat, useful for Discord modulation
  - Prevention of malicious content being shared on social media

INFERENCE

“I don't understand why people can't just act normal these days. Sure, everyone has their quirks, but it's getting ridiculous. Like, not every gay guy needs to be so flamboyant and over-the-top about everything. Meanwhile, at work, my coworker keeps complaining about the cafeteria food — it’s not even that bad. And then there’s the neighbors who keep leaving their trash cans out after pickup. People really need to get their act together.”

<0.5s

```
================================================================================
text: I don't understand why people can't just act normal these days.
Category: neutral
Stereotype Type: gender
================================================================================
text: Sure, everyone has their quirks, but it's getting ridiculous.
Category: neutral
Stereotype Type: profession
================================================================================
text: Like, not every gay guy needs to be so flamboyant and over-the-top about everything.
Category: stereotype
Stereotype Type: gender
================================================================================
text: Meanwhile, at work, my coworker keeps complaining about the cafeteria food—it’s not even that bad.
Category: neutral
Stereotype Type: profession
================================================================================
text: And then there’s the neighbors who keep leaving their trash cans out after pickup.
Category: neutral
Stereotype Type: profession
================================================================================
text: People really need to get their act together.
Category: neutral
Stereotype Type: profession
```

RISKS AND LIMITATIONS

- Limitations
  - Our model only works for sentences of similar sizes
  - Only works for these specific stereotypes
- Area of improvement
  - For bigger more complex pieces of text, taking the DL approach might be better
- Risk Verticals
  - Good: Performance, Explainability, Sustainability
  - Area of Improvement: Fairness (Bias)

BIAS ASSESMENT

```
text: Males usually think that the chef is talented.
Category: ['stereotype'], Probabilities: [[0.33614628 0.66385372]]
Stereotype Type: ['profession'], Probabilities: [[1.07562355e-02 1.80798294e-06 1.73505622e-03 9.87460682e-01
  5.61613055e-06 4.06017362e-05]]
[LightGBM] [Warning] min_data_in_leaf is set=2, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=2
[LightGBM] [Warning] min_data_in_leaf is set=2, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=2
==========================================================================================
text: Females usually think that the chef is talented.
Category: ['stereotype'], Probabilities: [[0.33145615 0.66854385]]
Stereotype Type: ['profession'], Probabilities: [[2.24267534e-02 2.49267064e-06 1.45551289e-03 9.76085153e-01
  9.30805575e-06 2.07802182e-05]]
```

```
text: Black's usually think that the chef is talented.
Category: ['stereotype'], Probabilities: [[0.15127715 0.84872285]]
Stereotype Type: ['profession'], Probabilities: [[1.65889435e-03 3.84202106e-06 9.45903403e-03 9.88580114e-01
  1.59078602e-04 1.39036604e-04]]
[LightGBM] [Warning] min_data_in_leaf is set=2, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=2
[LightGBM] [Warning] min_data_in_leaf is set=2, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=2
==========================================================================================
text: Asians often assume that the nurse is dedicated.
Category: ['neutral'], Probabilities: [[0.70386117 0.29613883]]
Stereotype Type: ['nationality'], Probabilities: [[1.52915770e-03 8.59569359e-06 6.25223406e-01 3.73088633e-01
  2.14520757e-05 1.28755803e-04]]
```

Thank you!