

# Training an NLP on Habitual Be A First Step Towards an AAVE-literate Language Learning Model

---

Morgan Goode | August 24, 2023



# Morgan Goode



## Experience

- 15+ years digital marketing/comms in the nonprofit sector

## Education

- BFA | Photography | Parsons School of Design | May 2008
- BA | Creative Nonfiction | Eugene Lang College | May 2008
- Data Science Bootcamp | Flatiron School | August 2023
- MA | Biography & Memoir | CUNY Graduate Center | May 2024

**Fun Fact:** Live Storyteller

[github.com/morgangoode](https://github.com/morgangoode) [linkedin.com/in/morgangoode](https://linkedin.com/in/morgangoode)

# Who be eating cookies?



Source: **Tense and Aspectual be in Child African American English**. Janice E. Jackson & Lisa Green

## Dialogue

Cookie Monster is sick and not eating cookies today. Elmo is eating cookies. Ernie only eats cookies on his birthday when his mom lets him. Cat has never had a cookie. Cats can't eat cookies.

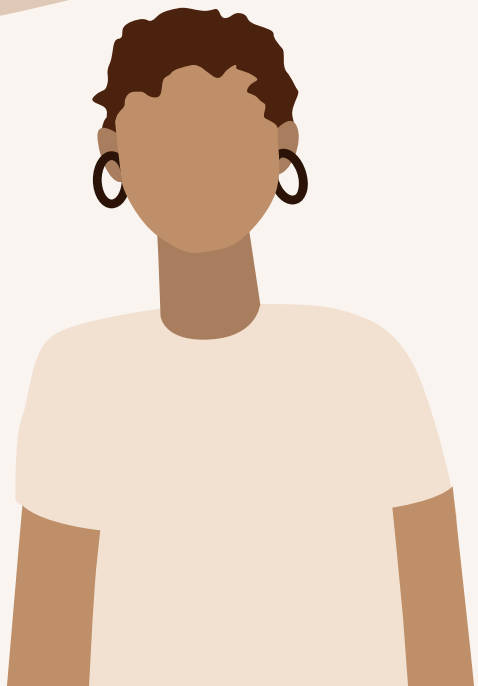
## Task Questions

1. Who be eating cookies?
2. Who is eating cookies?
3. Who eats cookies?
4. Who don't be eating cookies?
5. Who doesn't eat cookies?
6. Who isn't eating cookies?

## Correct Responses

1. Cookie Monster
2. Elmo
3. Cookie Monster, Ernie, Elmo
4. Cat, Ernie, Elmo
5. Cat, Ernie
6. Cookie Monster, Ernie, Cat

**We can  
predict  
habitual be  
with 90%  
accuracy**





# Agenda

---

**01**

Objectives +  
Methodology

**02**

Data Overview

**03**

Final Model

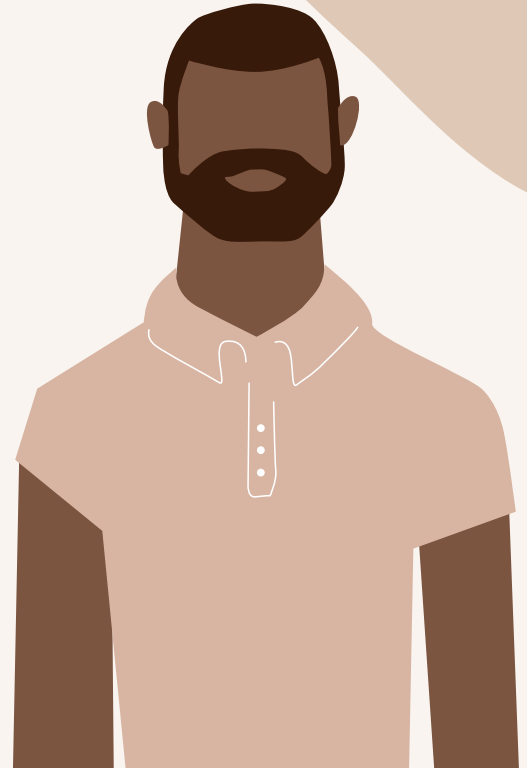
**04**

Conclusion +  
Next Steps



01

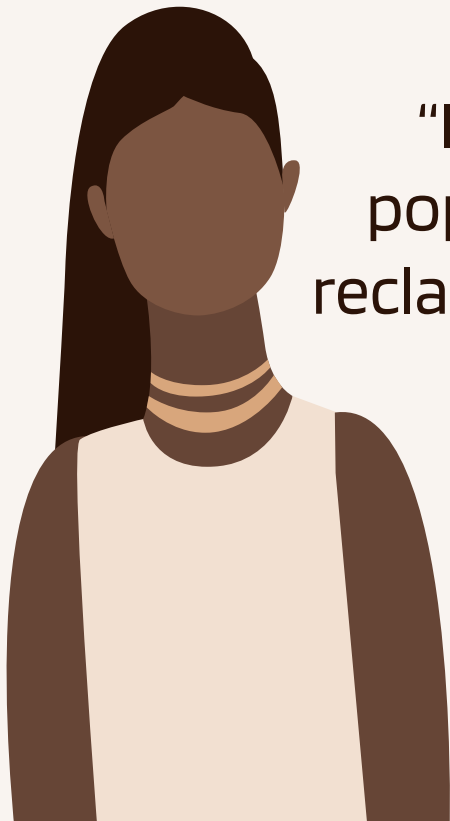
# Objectives + Methodology



A close-up photograph of several hands of different skin tones (Black, Brown, and White) gently cupping a small, light-colored object in the center. The hands are positioned in a way that suggests care and support. The background is softly blurred, focusing attention on the hands and the object they are holding.

## Objective: Inclusion + Accuracy Sans Bias

In other words: a probabilistic algorithm for identifying speech patterns that does **not** encode anti-Black bias through negative word associations.



"If we filter out the discourse of marginalized populations, we fail to provide training data that reclaims slurs and otherwise describes marginalized identities in a positive light."

— **On the Dangers of Stochastic Parrots:  
Can Language Models Be Too Big?**

Source: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?  
Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell  
<https://dl.acm.org/doi/pdf/10.1145/3442188.3445922>



A background image showing several hands of different skin tones (dark brown, light brown, and white) stacked together, palms up, in a gesture of support or unity. The hands are positioned in the lower half of the frame, with the fingers spread out. The background is slightly blurred, focusing attention on the hands.

# Methodology

---

- Hand curated dataset
- Including 'stop words'
- No filtering of obscenities or slurs
- Data documentation
- Ongoing corpus edits/expansion to ensure balance and mitigate bias

02

# Data Overview



# Data Overview

---



## 2K Corpus

Manually Compiled +  
Tagged



## Native AAVE speakers

In all habitual be records  
+ most present be  
records

# Data Limitations + Room for Improvement

---



## Team of One

Collaboration with  
stakeholders is a must



## No Sentiment Analysis (Yet!)

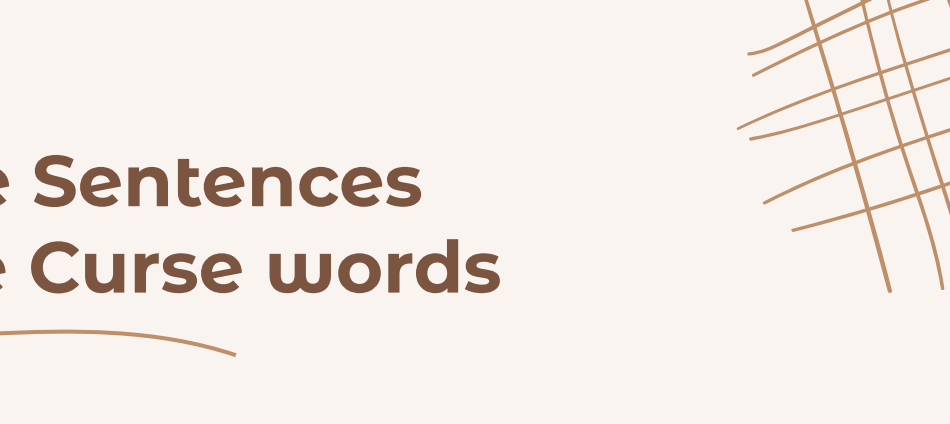
Imbalanced sentiment could  
contribute to bias



## Speaker Diversity

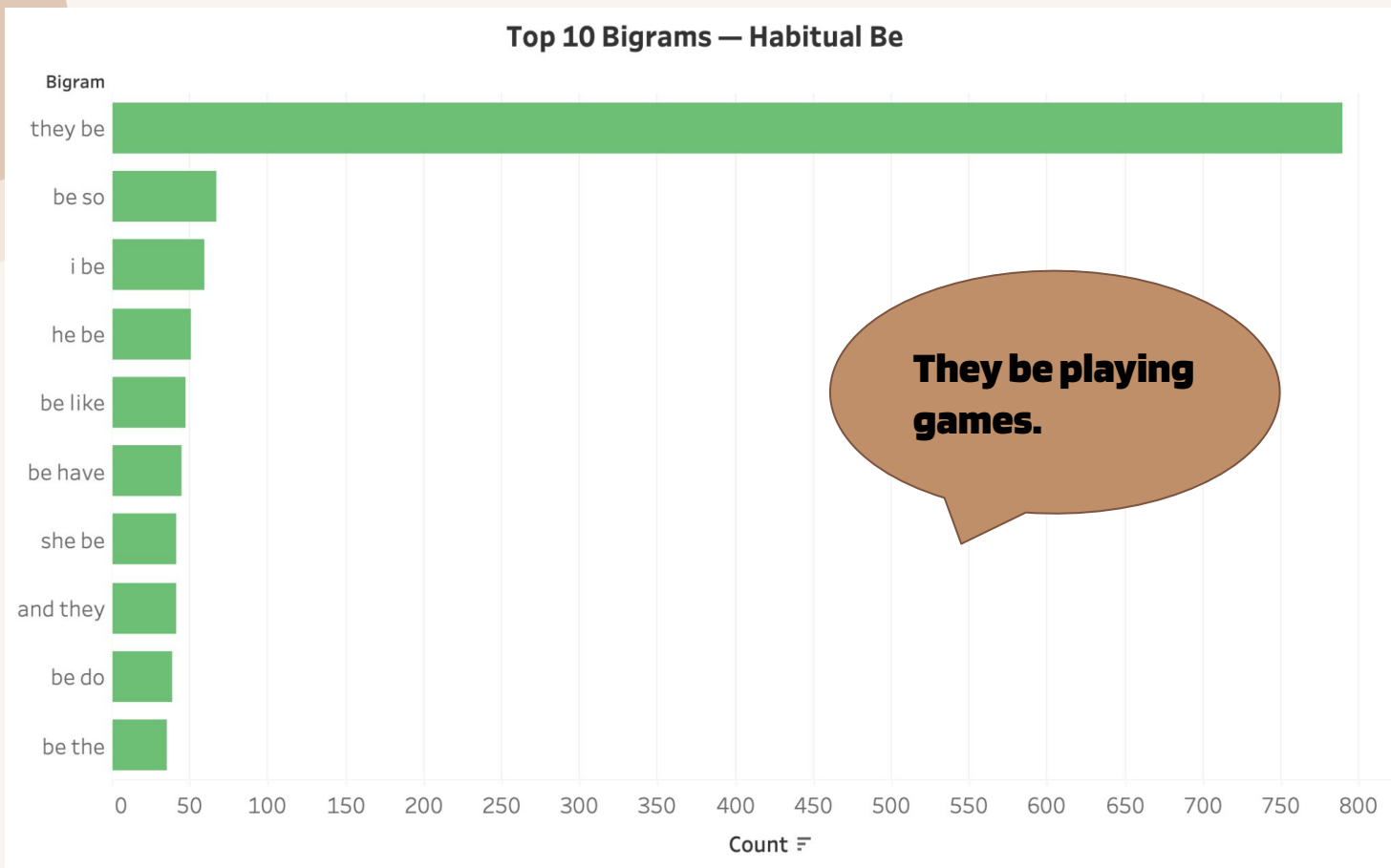
Curate and confirm balanced  
representation across age, class,  
gender, sexual orientation, location,  
and other demographics

## Habitual Be Word Cloud

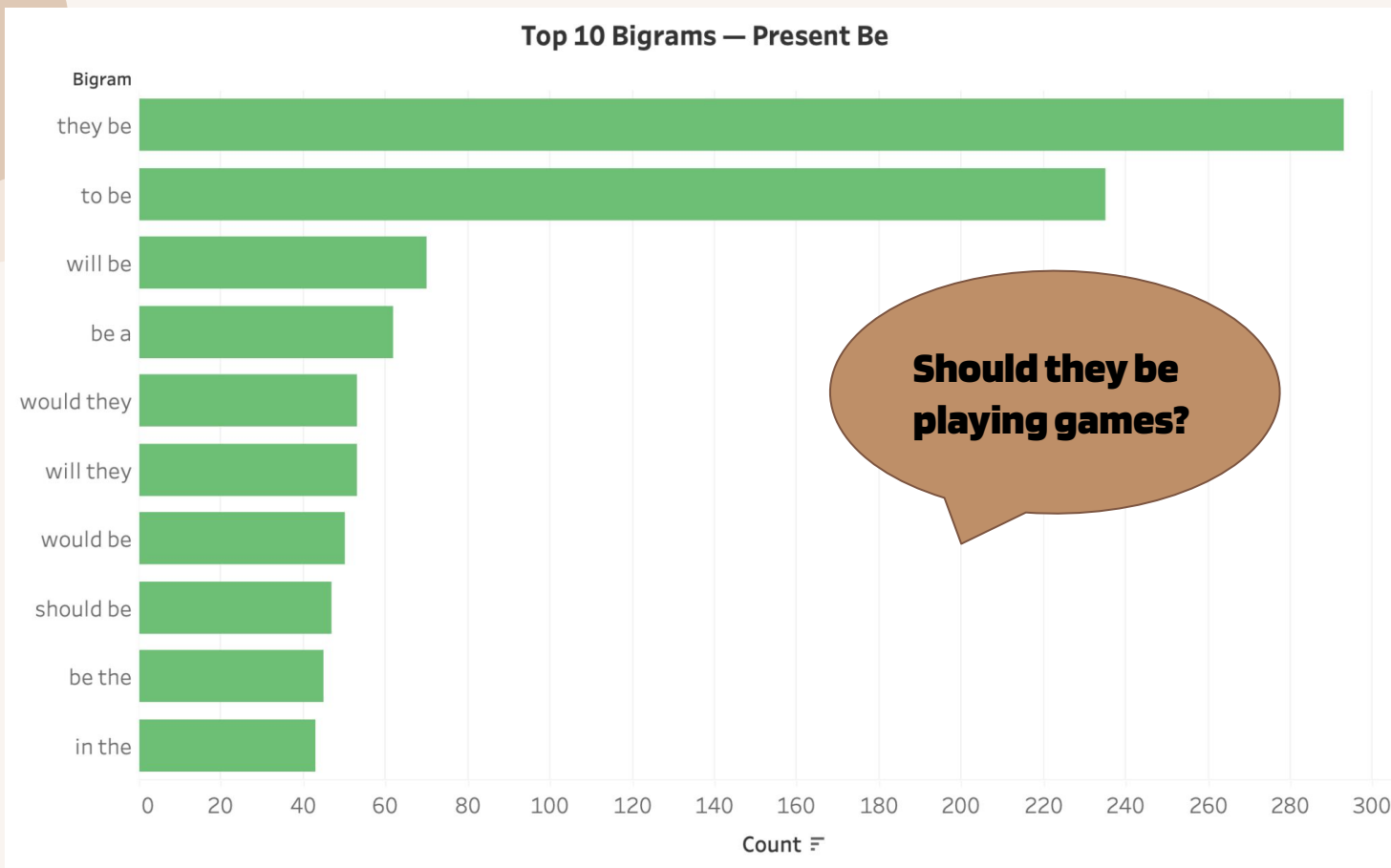


## Present Be Word Cloud

# Bigrams are Key



# Bigrams are Key



03

# Final Model





# Multinomial Naive Bayes

---



|           | Training | Test |
|-----------|----------|------|
| Accuracy  | 93%      | 90%  |
| Recall    | 95%      | 91%  |
| Precision | 91%      | 89%  |
| F1        | 93%      | 90%  |

04

# Conclusion + Next Steps





## Next Steps

---

**Edit +  
Expand  
Corpus**


In collaboration with other  
AAVE speakers + linguists

**Model  
Iterations**

support vector machines +  
neural networks

**Train  
models**

Incorporate more features  
of AAVE





# Thank you!

---

Any questions? Drop  
them in the Q & A!

**Morgan Goode**

[github.com/morgangoode](https://github.com/morgangoode)

[linkedin.com/in/morgangoode](https://linkedin.com/in/morgangoode)



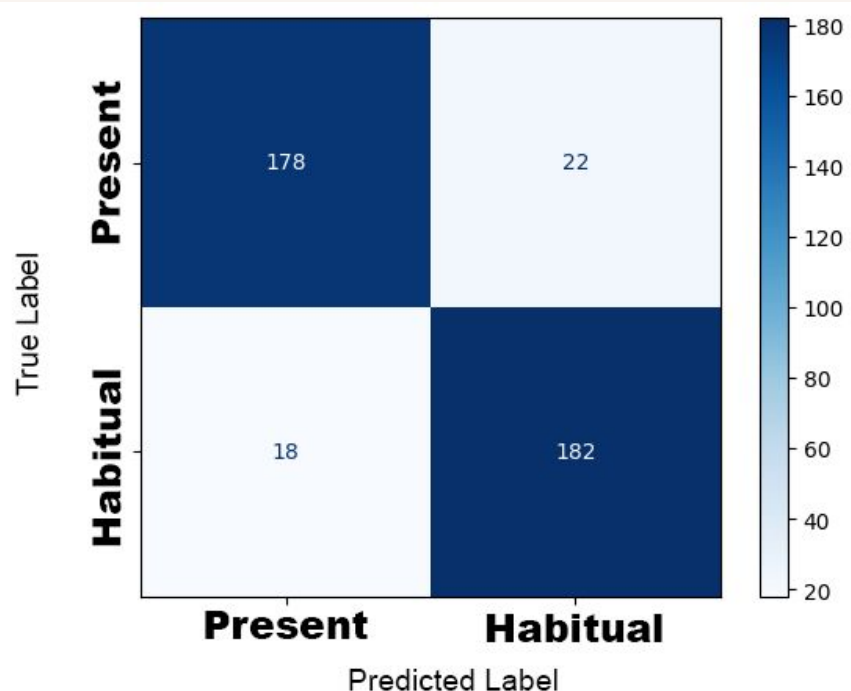
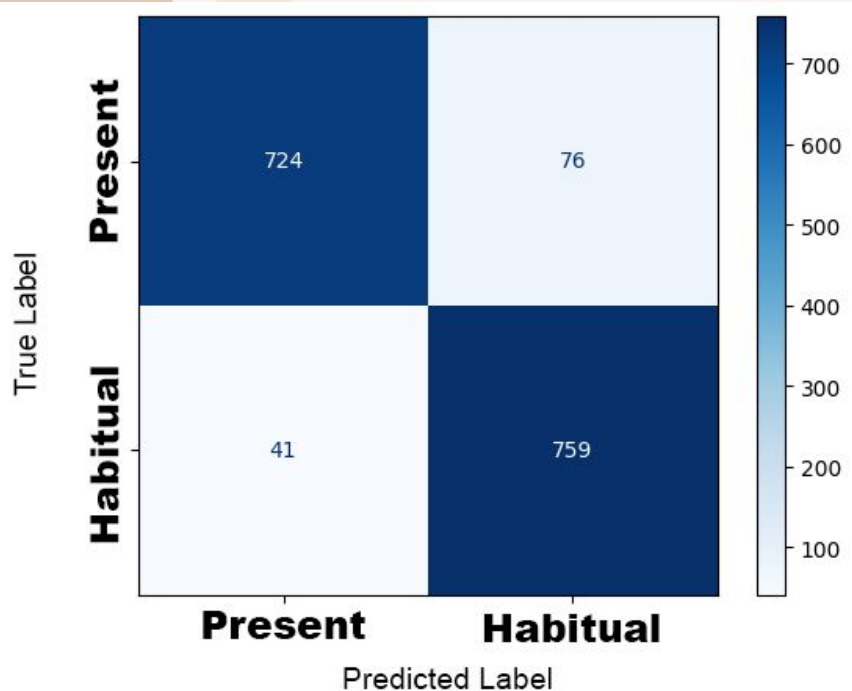


# Appendix

# Confusion Matrices

Train

Test



# Sources & Further Reading

- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. ***On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*** 🦜 In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Emily M. Bender and Batya Friedman. 2018. ***Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science.*** Transactions of the Association for Computational Linguistics, 6:587–604. <https://aclanthology.org/Q18-1041>
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford. 2018. ***Datasheets for Datasets*** <https://arxiv.org/abs/1803.09010>
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. ***RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models.*** In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3356–3369, Online. Association for Computational Linguistics.
- Janice E. Jackson, Lisa Green. 2005. ***Tense and Aspectual be in Child African American English.*** In: Verkuyl, H.J., de Swart, H., van Hout, A. (eds) Perspectives on Aspect. Studies in Theoretical Psycholinguistics, vol 32. Springer, Dordrecht. [https://doi.org/10.1007/1-4020-3232-3\\_13](https://doi.org/10.1007/1-4020-3232-3_13)
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. ***Model Cards for Model Reporting.*** In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT\* '19). Association for Computing Machinery, New York, NY, USA, 220–229. <https://doi.org/10.1145/3287560.3287596>