

Tori Stiegman ‘23, Data Science Capstone Project

Background and Research Questions

Gender-inclusive language, or spoken or written word that does not discriminate against a particular sex, social gender or gender identity, and does not perpetuate gender stereotypes (United Nations, 2017), is one of those areas, especially regarding female reproductive organs. While many people with uteruses identify as cis-gendered women, not all do. Yet, when discussing menstruation, people often discuss it as a problem “only women” face, reinforcing the false equivalency between women and people who menstruate. However, people can still be considered a “woman” even if they do not menstruate, and people who menstruate are not necessarily women. Discussions surrounding menstruation targeted mainly toward ciswomen can exclude people from something happening to their bodies, be potential sources of gender dysphoria and discomfort, and even become sources of discrimination and safety concerns (Anagnostou, 2021). Thus, using gender-inclusive language while talking about menstruation will ensure that everyone can access the information they need safely and help American society work towards dismantling gender biases and stigmas surrounding menstruation.

This study utilizes two multinomial naive bayes classifiers and two classification trees to categorize tweets as gender-inclusive or exclusive or neither in order to answer these particular research questions:

1. What kind of conversations are Twitter users having about menstruation?
2. Do Twitter users use gender-inclusive language while talking about menstruation?
3. Can a classifier be used to determine whether or not a tweet is inclusive?

The driving hypothesis for this research is that **while some Twitter users will likely use gender-inclusive language, more users will use gender-exclusive language or neither gender-inclusive nor exclusive language when talking about menstruation.**

Data

Description

The Twitter API was used to collect tweets from November 10, 2020, through November 10, 2022 that contained: “menstrual cycle,” “tracking your period,” “period tracking,” “track your menstrual cycle,” “period tracker,” or “menstruation.” After dropping the second instance of duplicate tweets, the final raw dataset contained 301,153 tweets. Each tweet had information about the tweet id, author id, the tweet’s text and date, the number of times the tweet was retweeted and the number of likes the tweet received.

This dataset was then split into three separate datasets, one that was used for **training (n = 280)**, one for **testing (n = 56)** and one that contained the **rest of the tweets (n = 300,817)**, using cluster sampling.

Cleaning

A new variable, **text_clean**, was added to each dataset containing the “cleaned” version of the tweet text that did not contain emojis, hashtags, mentions, links, punctuations and non-alphanumeric characters, was all lowercase and stemmed.

	text	label	tweet_id	author_id	date	retweet_count	reply_count	like_count	text_clean
0	These resources are awesome. Today is #WorldMenstrualHygieneDay - approximately half of the population experiences menstruation, and more than 2/5 of them report struggling to afford period products. There's a sharp racial divide and the pandemic has only made it worse. https://t.co/3bHTPA7Jd	inclusive			2021-05-28	0	0	0	resourc awesom today approxim half popul experi menstruat 2 5 them report struggl afford period product sharp racial divid pandem made wors
1	Wherein women were not "weakened" during these days and were capable of all activities. \n\nThis worked well and took the stigma out of menstruation - think back when we were kids and how hush we were about "Aunt Flo" (or cultural equivalent) vs. how open we are now	exclusive			2022-02-26	0	1	0	wherein women weaken day capabl activ work well took stigma menstruat think back kid hush aunt flo cultur equival vs open

Methods

First, the training and testing data was labeled as either gender-inclusive, exclusive or neither based on specific guidelines that operationalized the meaning of “gender-inclusive” and “gender-exclusive.”

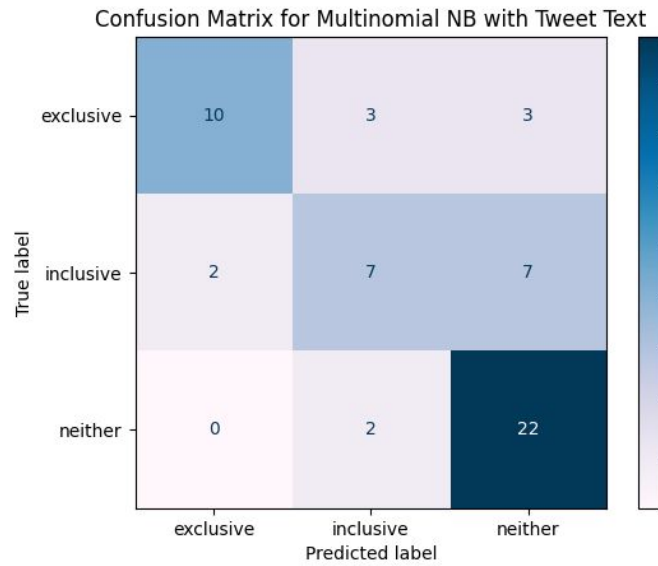
After labeling, there were 73 gender-exclusive tweets, 64 gender-inclusive tweets and 143 tweets that were neither in the training dataset (n = 280). In the test dataset (n = 56), there were 16 gender-exclusive tweets, 16 gender-inclusive tweets and 24 tweets that were neither.

The cleaned text and the labels from the training dataset (n = 280) were then fed into two multinomial Naive Bayes classification models and two classification trees. The accuracies, sensitivities and specificities of each model were then used to determine the best model, the Classification Tree with Text model, which was then used to label the larger tweet corpus.

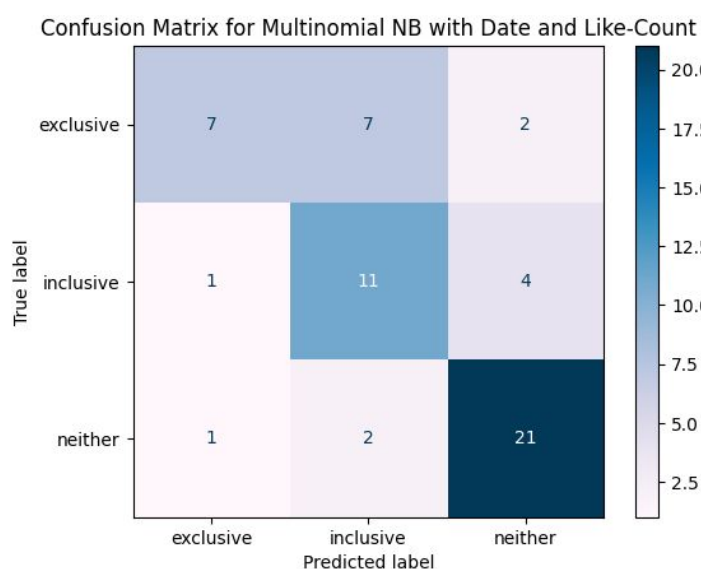
Classifiers

Naive Bayes with Text

- Accuracy: 0.696
- Sensitivity:
 - Exclusive: 0.625
 - Inclusive: 0.4375
 - Neither: 0.917
- Specificity: 0.952
 - Exclusive: 0.952
 - Inclusive: 0.89
 - Neither: 0.762



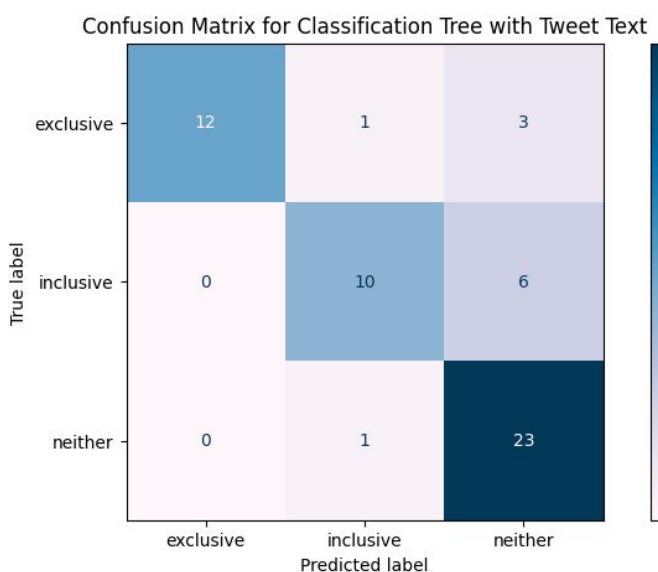
Naive Bayes with Text, Date, Likes



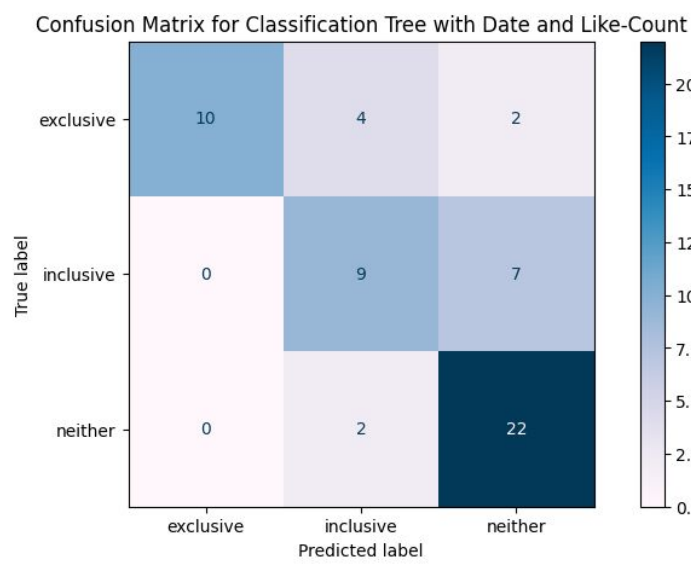
- Accuracy: 0.696
- Sensitivity:
 - Exclusive: 0.4375
 - Inclusive: 0.6875**
 - Neither: 0.875
- Specificity:
 - Exclusive: 0.952
 - Inclusive: 0.816
 - Neither: 0.842**

Classification Tree with Text

- Accuracy: 0.804**
- Sensitivity:
 - Exclusive: 0.75**
 - Inclusive: 0.625
 - Neither: 0.958**
- Specificity:
 - Exclusive: 1.0**
 - Inclusive: 0.952**
 - Neither: 0.780



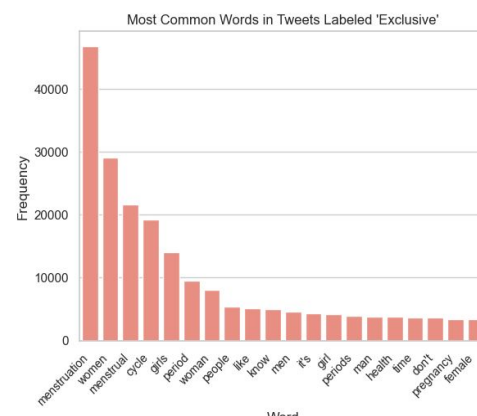
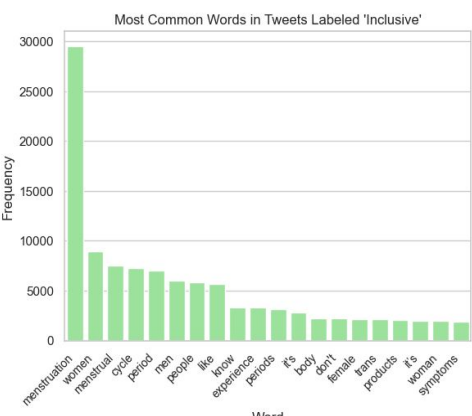
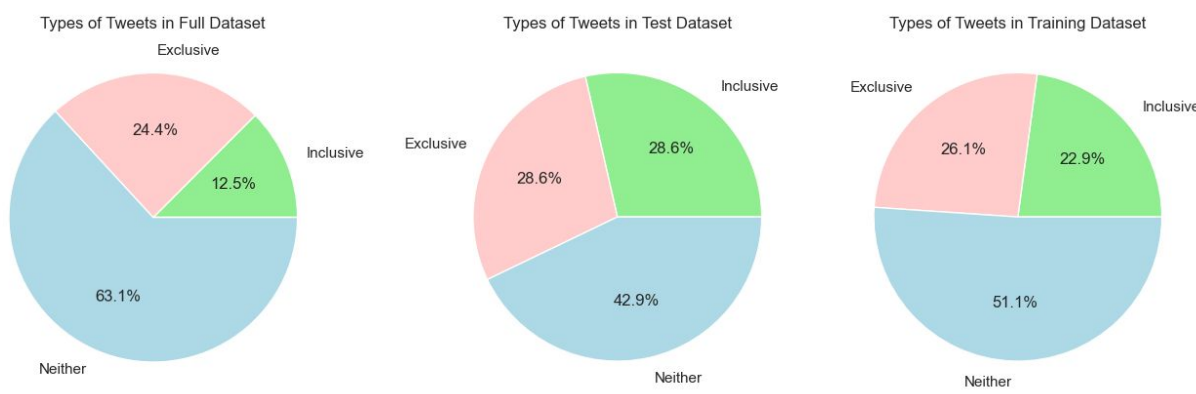
Classification Tree with Text, Date, Likes



- Accuracy: 0.732
- Sensitivity:
 - Exclusive: 0.625
 - Inclusive: 0.5625
 - Neither: 0.917
- Specificity:
 - Exclusive: 1.0**
 - Inclusive: 0.870
 - Neither: 0.780

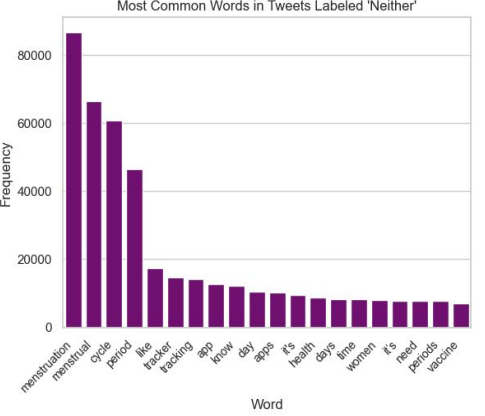
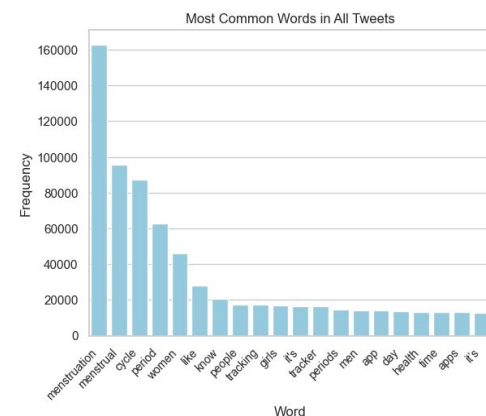
Results and Conclusions

When the chosen classification tree classifier was applied to the full corpus of the 300,817 tweets, 12.5% or 37,590 tweets used gender-inclusive language and 24.4.6% or 73,266 tweets used gender-exclusive language.



The types of words included in gender-inclusive and gender-exclusive tweets are not inherently different. However, gender-exclusive tweets tend to include many more instances of “girl” and “female,” whereas gender-inclusive tweets tend to contain more examples of the words “trans” and “sex.”

Gender-inclusive tweets also included many instances of “talk,” “think,” and “need.” On the other hand, gender-exclusive tweets include many instances of “hygiene,” “pregnancy” and “pain.”



These discrepancies could point towards these two separate groups having different agendas and topics of interest that all fall under the umbrella of “menstruation.” However, more work would need to be done to clarify and solidify the relationship between the inclusivity of language and the topic of discussion.

References

“Gender-inclusive communication,” *United Nations*, 2017. [Online]. Available: <https://www.un.org/en/gender-inclusive-language/#:~:text=Using%20gender%2Dinclusive%20language%20means,does%20not%20perpetuate%20gender%20stereotypes>. [Accessed: 19-Dec-2022].

J. Anagnostou, “How to talk about periods in a more inclusive way,” *Moxie*, 31-Mar-2021. [Online]. Available: <https://moxie.com.au/blogs/the-regular/how-to-talk-about-periods-in-a-more-inclusive-way#:~:te xt=Instead%2C%20you%20could%20use%20language,%22menstrual%20cups%22%2C%20etc>. [Accessed: 19-Dec-2022].