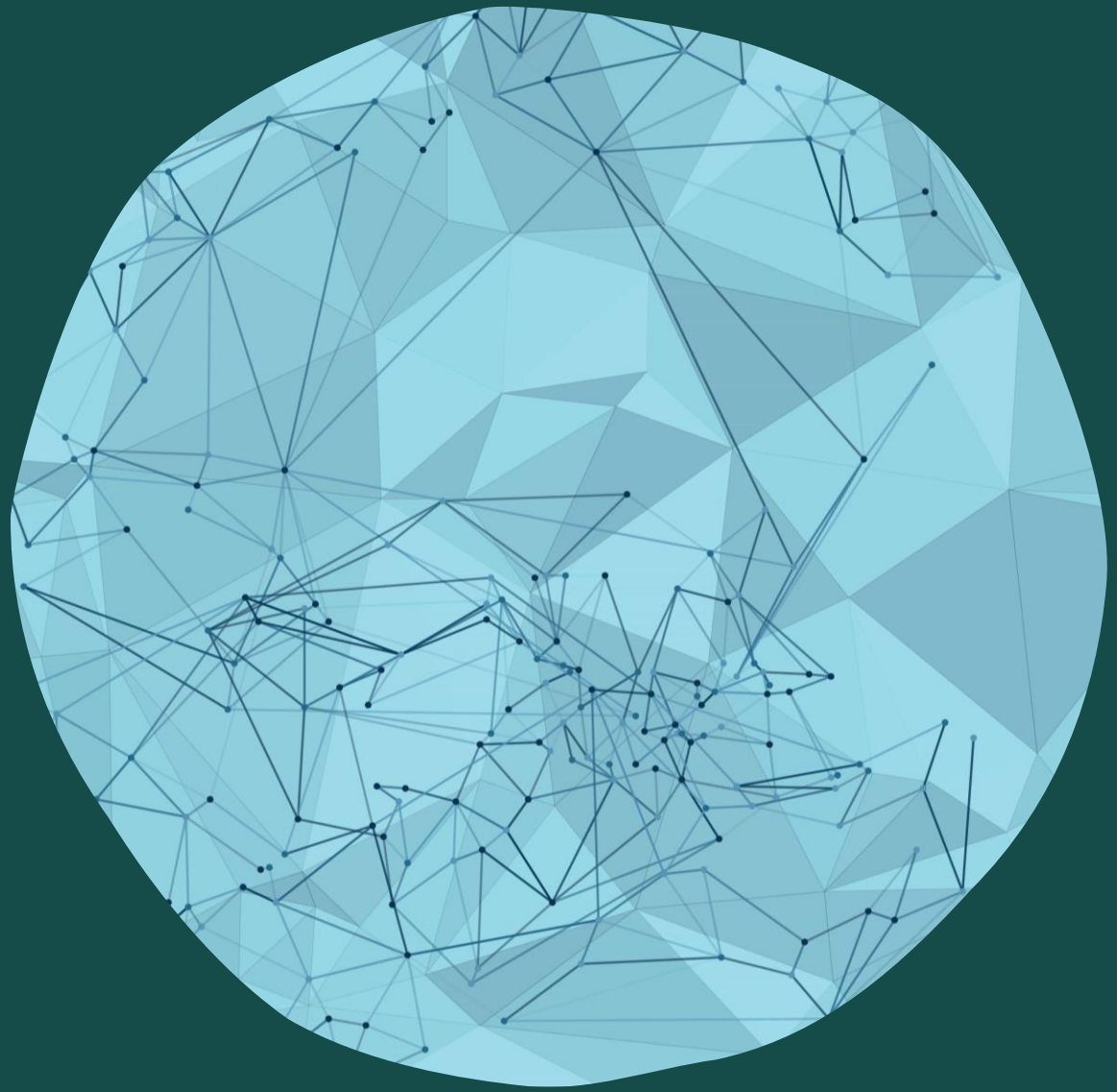


AVOIDING HARMS IN CLASSIFICATION & SUMMARY

Members

1.Huynh Hoang Vu - 20520864

2.Le The Viet - 20520093

3.Tran Huu Khoa - 20520222

4.Nguyen Duy Dat - 20520435

Tuesday, October 25th 2022

4.10 Avoiding Harms in Classification

a) Representational harms: caused by a system that demeans a social group (perpetuating negative stereotypes about them)

4.10 Avoiding Harms in Classification

[200 sentiment analysis systems on pairs of sentences]

African American first name (like Shaniqua)
European American first name (like Stephanie)

+ Most systems assigned lower sentiment & more negative emotion to sentences with above cases

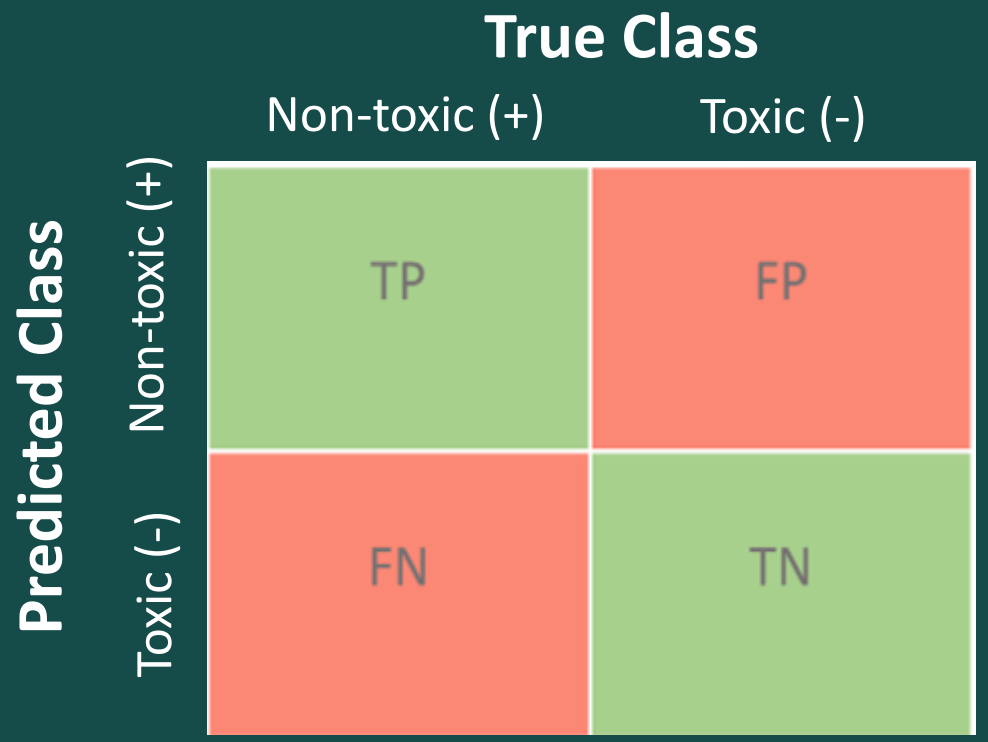
=> Reflecting and perpetuating stereotypes that African Americans with negative emotions

4.10 Avoiding Harms in Classification

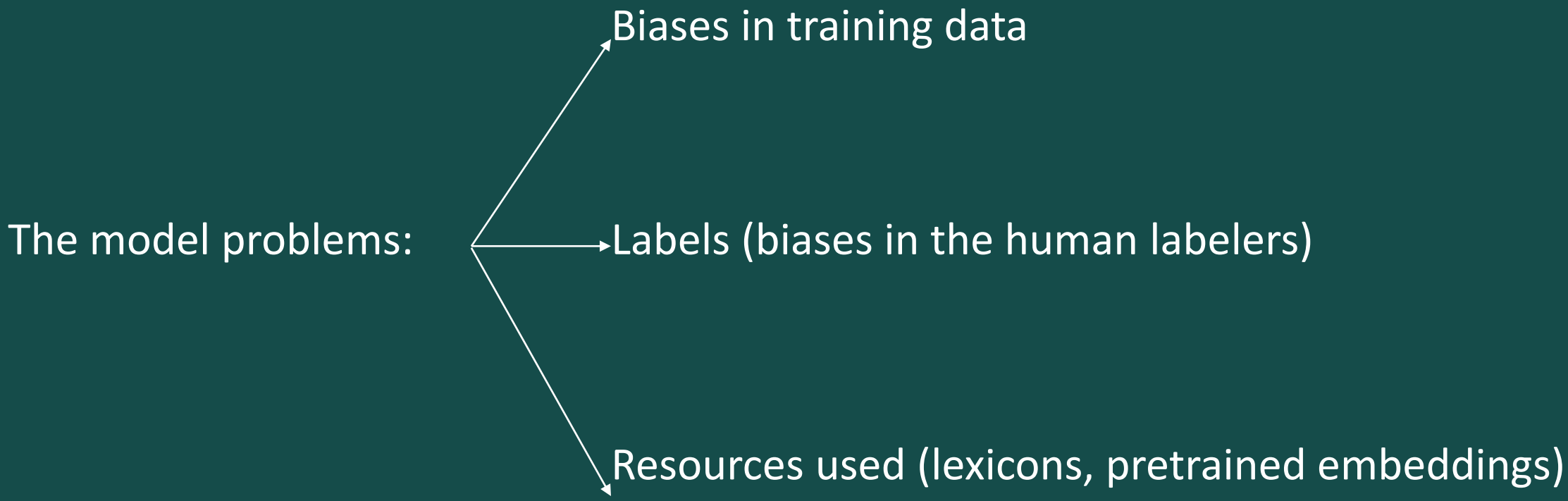
b) Toxicity detection: detecting hate speech, abuse, harassment, or other kinds of toxic language.

4.10 Avoiding Harms in Classification

Harms: toxic sentences -> non-toxic sentences



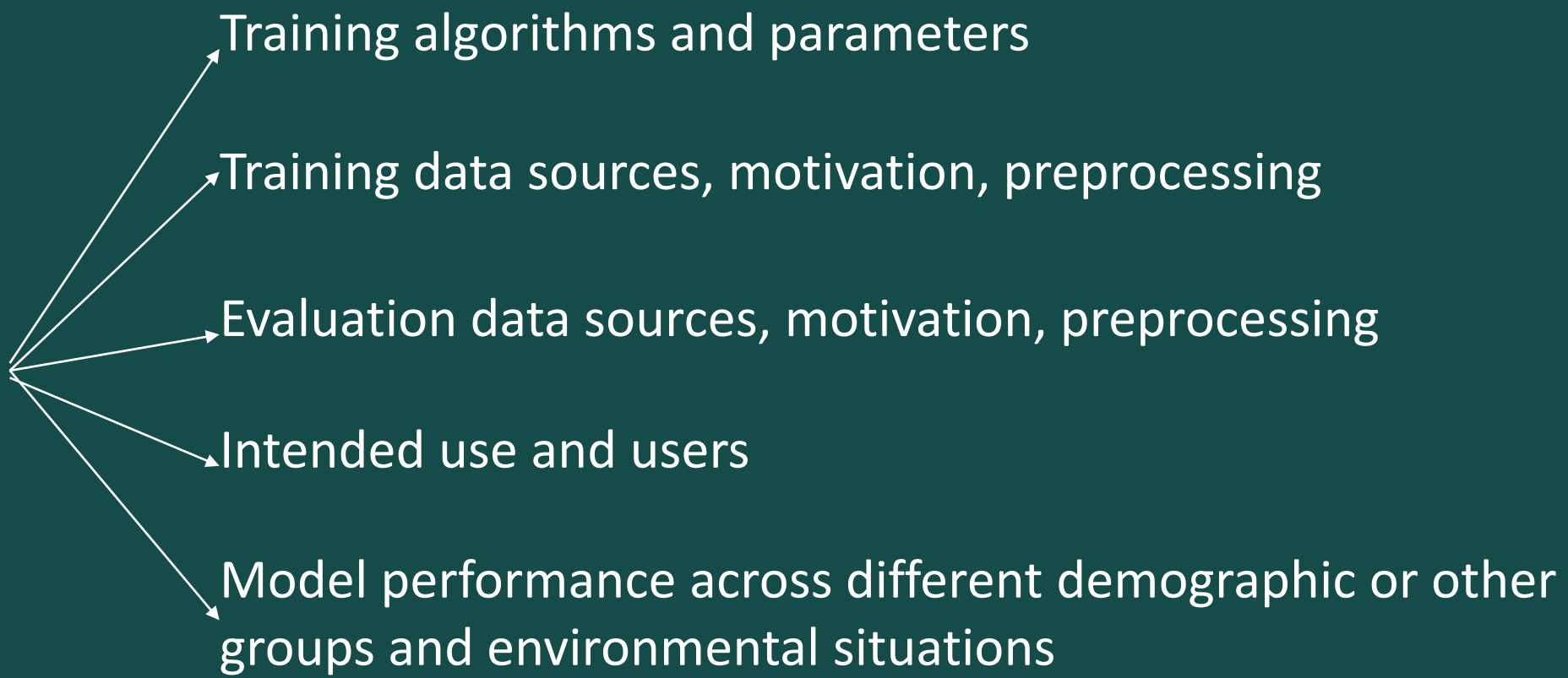
4.10 Avoiding Harms in Classification



=> The mitigation of biases (carefully considering the training data sources) is an important of research, we currently don't have general solutions.

4.10 Avoiding Harms in Classification

c) Model card:



4.11 Summary

This chapter introduced "Naive Bayes" model for classification and applied it to the text categorization task of sentiment analysis.

+ Many language processing tasks can be viewed as tasks of classification

+ Text categorization: sentiment analysis, spam detection, …

+ Sentiment analysis classifies a text as reflecting (+) or (-) orientation

+ Naïve Bayes is a generative model that makes the bag of words assumption and the conditional independence assumption.

+ Naïve Bayes with binarized features work better for many text classification.

4.11 Summary

+ Classifiers are evaluated based on “precision” and “recall”

+ Classifier are trained using distinct train/dev/test sets, including “cross-validation” in training set

+ Statistical tests should be used to determine whether we can be confident that one version of a classifier is better than another.

+ Designers of classifiers should consider harms that may be caused by the model, including its training data, report model characteristics in “model card”