Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

## **Supervised Topic Models**

Advanced Machine Learning for NLP
Jordan Boyd-Graber
MULTILINGUAL APPROACH

- We have documents in multiple languages
- They are annotated with the same **continuous response**
  - Rating of a product
  - Movie review
  - Percentage of people who retweet a tweet
  - Percentage of people consider a comment "extreme"
- Can learning a model in multiple languages at once help?

- We have documents in multiple languages
- They are annotated with the same **continuous response**
  - Rating of a product
  - Movie review
  - Percentage of people who retweet a tweet
  - Percentage of people consider a comment "extreme"
- Can learning a model in multiple languages at once help?

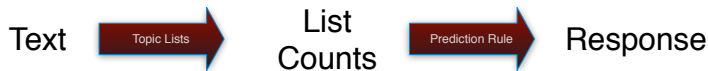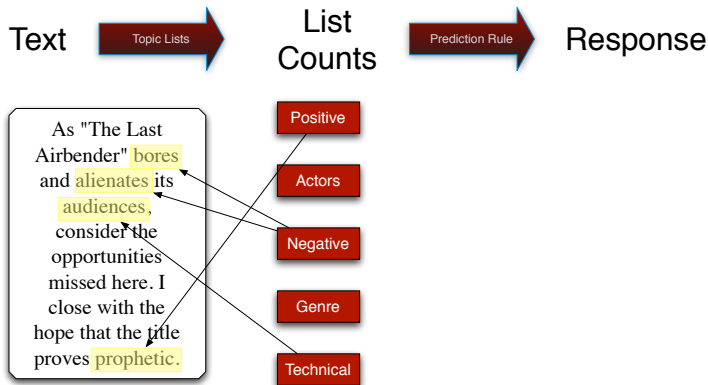For languages where you don't have many resources, **yes!**

**Conceptual Approach**

Text → *Topic Lists* → List Counts → *Prediction Rule* → Response

## Conceptual Approach

Text → Topic Lists → List Counts → Prediction Rule → Response

> As "The Last Airbender" bores and alienates its audiences, consider the opportunities missed here. I close with the hope that the title proves prophetic.

## Conceptual Approach

## Conceptual Approach



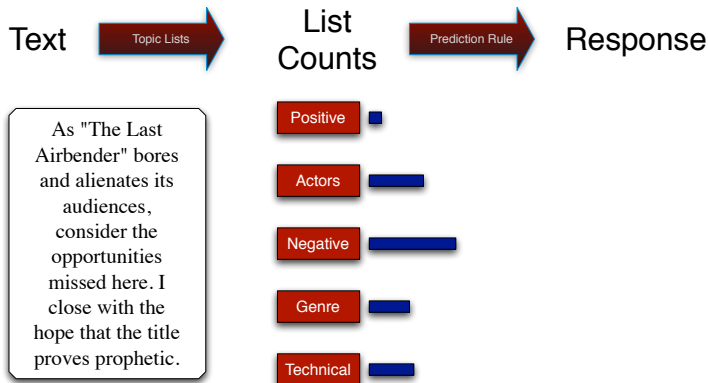Text → Topic Lists → List Counts → Prediction Rule → Response

> As "The Last Airbender" bores and alienates its audiences, consider the opportunities missed here. I close with the hope that the title proves prophetic.

Positive ▪
Actors ▬
Negative ▬▬
Genre ▬
Technical ▬

## Conceptual Approach



Text → Topic Lists → List Counts → Prediction Rule → Prediction

As "The Last Airbender" bores and alienates its audiences, consider the opportunities missed here. I close with the hope that the title proves prophetic.
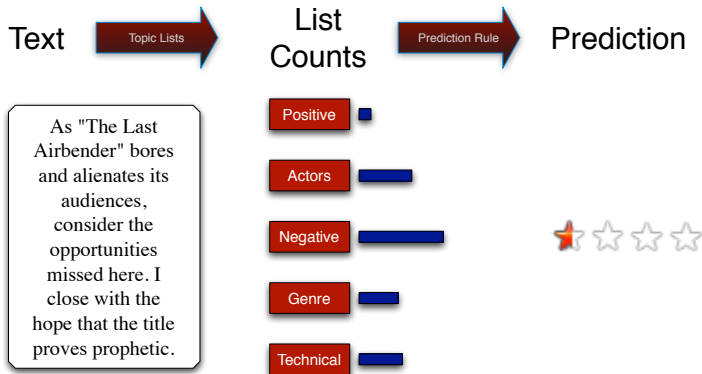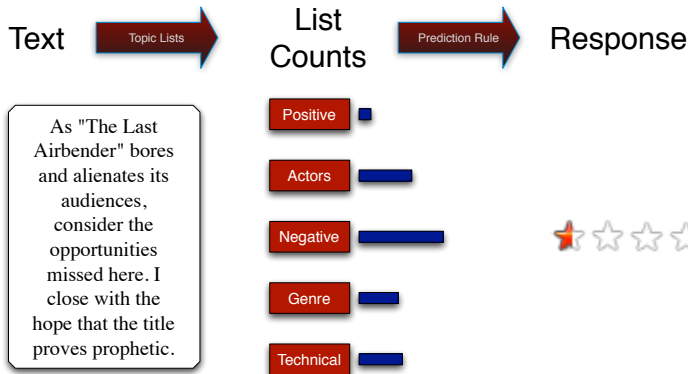
Positive

Actors

Negative

Genre

Technical

Similar to social science methodology LIWC **?**

Text → Topic Lists → List Counts → Prediction Rule → Response

As "The Last Airbender" bores and alienates its audiences, consider the opportunities missed here. I close with the hope that the title proves prophetic.

Positive

Actors

Negative

Genre

Technical

- **Assumption:** We can create list counts from documents in any language
- **Observation:** Once we have list counts, underlying language doesn't matter

Text → Topic Lists → List Counts → Prediction Rule → Response

> As "The Last Airbender" bores and alienates its audiences, consider the opportunities missed here. I close with the hope that the title proves prophetic.

**?** Positive

**?** Angles

**?** Negative

**?** Critics

**?** Technical

★ ☆ ☆ ☆
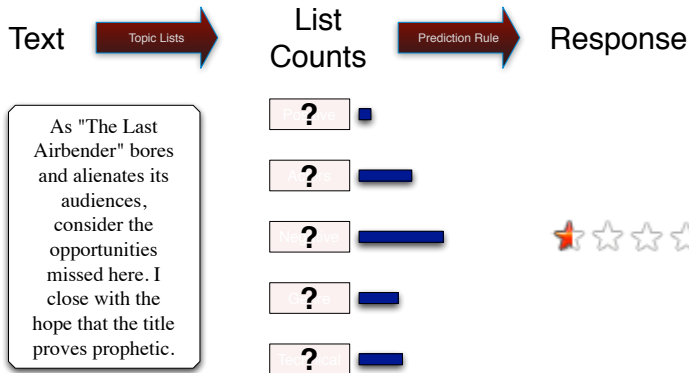
What if we don't know the lists?

- How do we learn the word lists?

- How do ensure that the word lists reflect sentiment?

- How do make the word lists make sense across languages?

- How do we learn the word lists?

- How do ensure that the word lists reflect sentiment?

- How do make the word lists make sense across languages?

- How do we learn the word lists?
  - Topic Models
- How do ensure that the word lists reflect sentiment?

- How do make the word lists make sense across languages?

- How do we learn the word lists?
  - Topic Models
- How do ensure that the word lists reflect sentiment?
  - Supervised Topic Models
- How do make the word lists make sense across languages?

- How do we learn the word lists?
  - Topic Models
- How do ensure that the word lists reflect sentiment?
  - Supervised Topic Models
- How do make the word lists make sense across languages?
  - Semantic Resources

**Why do this?**

- Topic models embed documents in low dimensional space
- These spaces are often useful for prediction
- But not designed for it!
- Can we use different objective functions to optimize embedding
- Understanding interplay between

- Supervised topic models
- Using multiple languages
- Hierarchical non-parametric models