


Pythons vs. Gophers, Who wins?



Presenter: Christopher McGowen



What are the differences
between users of Python and
Golang?

Process:



```
graph LR; A[Data Gathering] --> B[Data Cleaning]; B --> C[Feature Engineering]; C --> D[Modeling]
```

Data Gathering

Data Cleaning

Feature Engineering

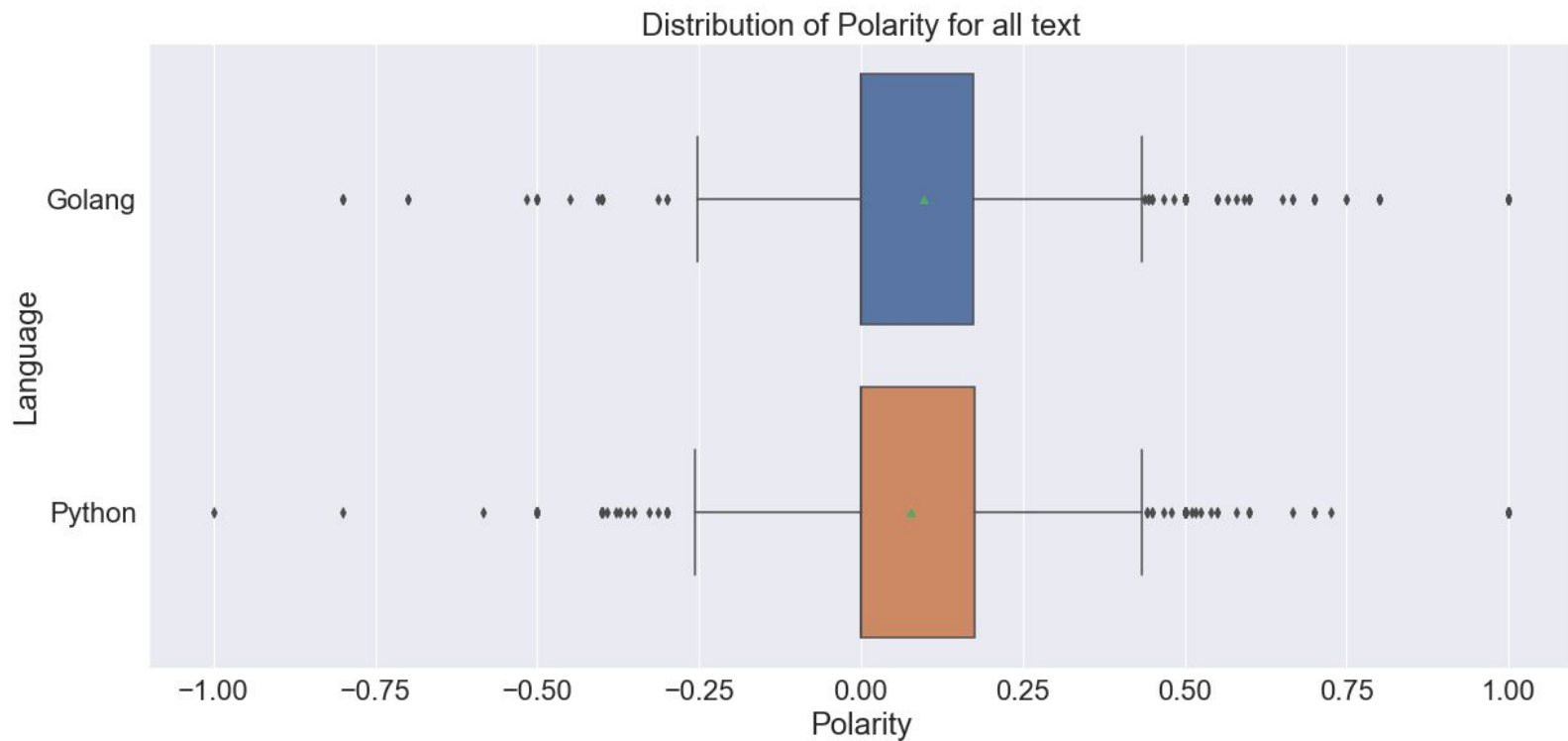
Modeling

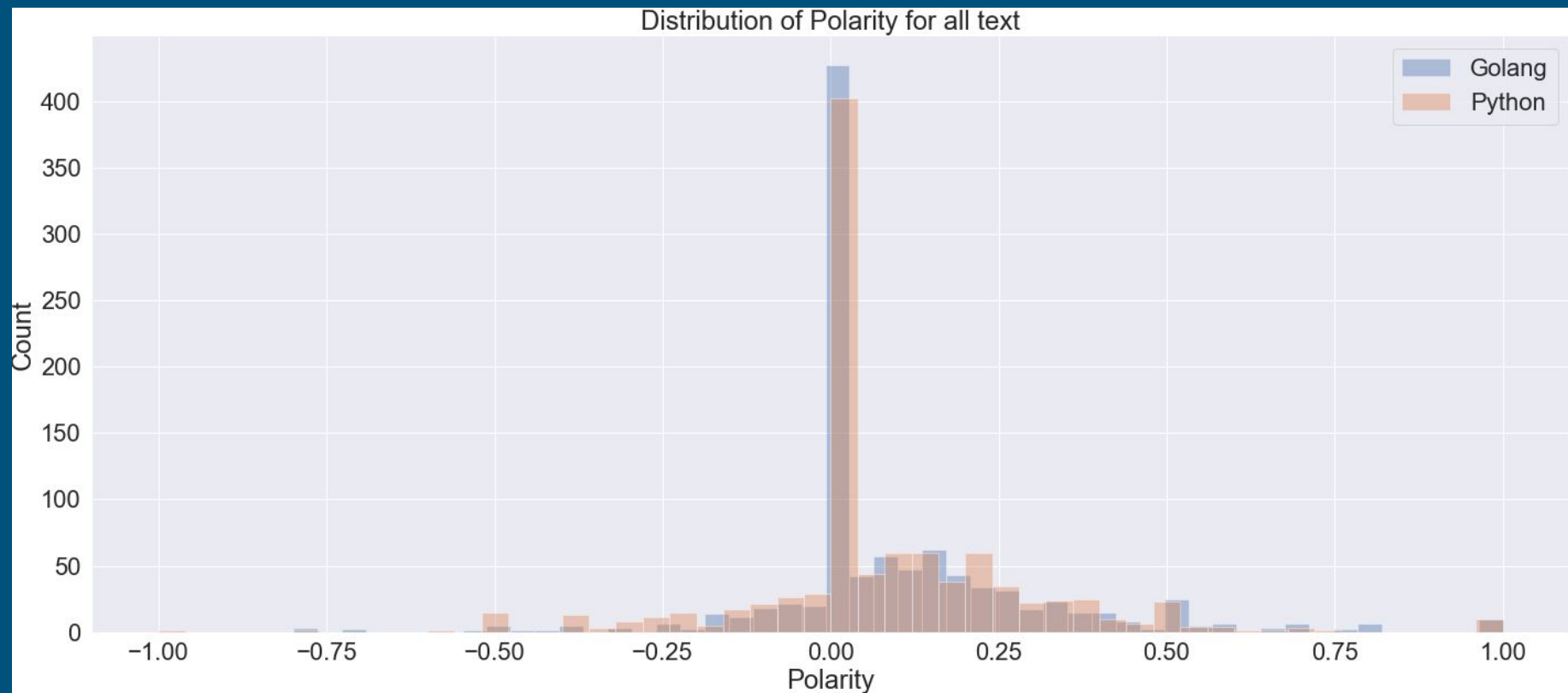
- Used Pushshift API to pull 1000 r/python and r/Golang posts (2000 total)
 - Only posts with ≥ 10 comments
 - Only captured title, body, and subreddit name

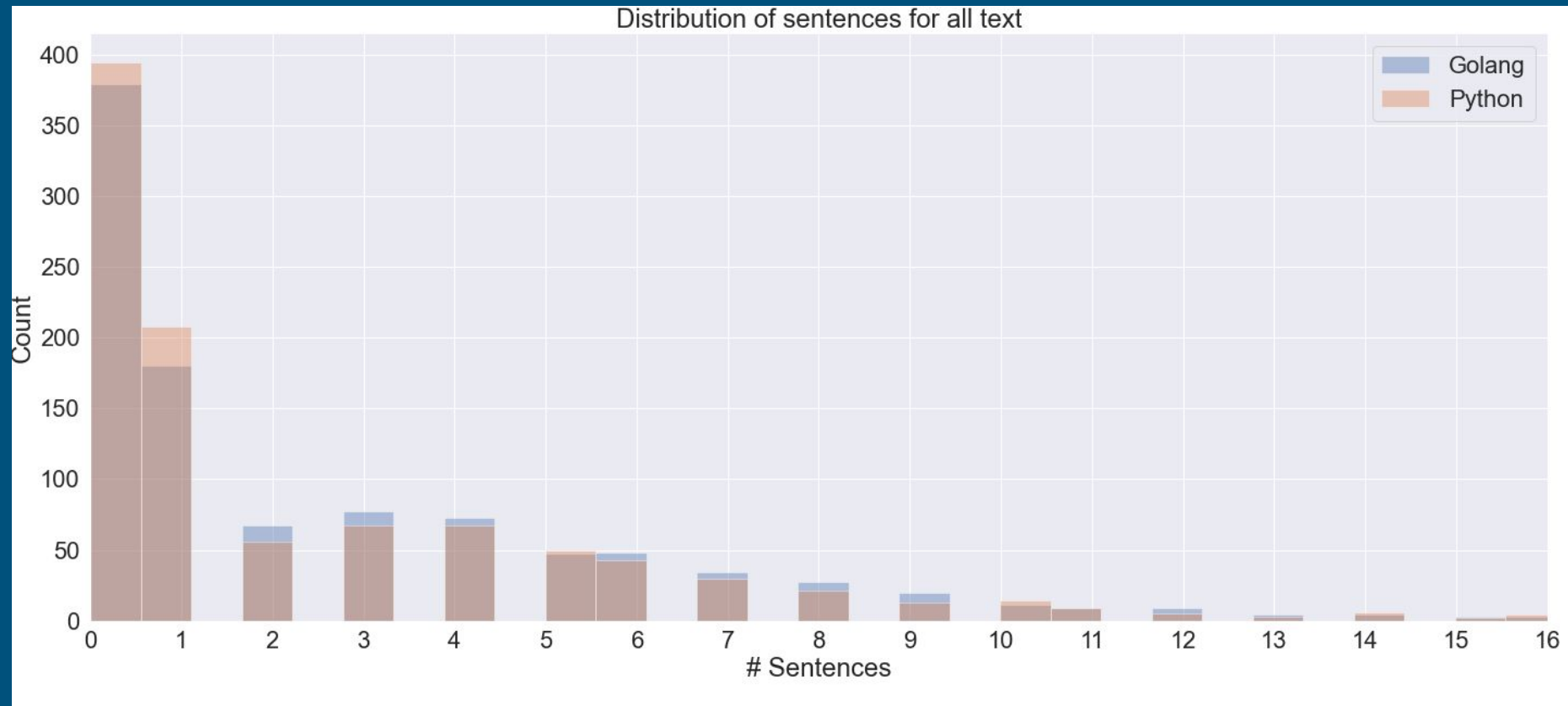
- Removed html tags with BeautifulSoup
- Removed non-alphabet characters
- Tokenized text into sentences, and words with TextBlob

Hypothesis: Posts will have different sentiments

- Created new features with TextBlob
 - Sentiment analysis (Polarity & Subjectivity)
 - Polarity = Score from -1 to 1, negative to positive sentiment
 - Subjectivity = Score from 0 to 1, objective to subjective
 - Word and Sentence counts

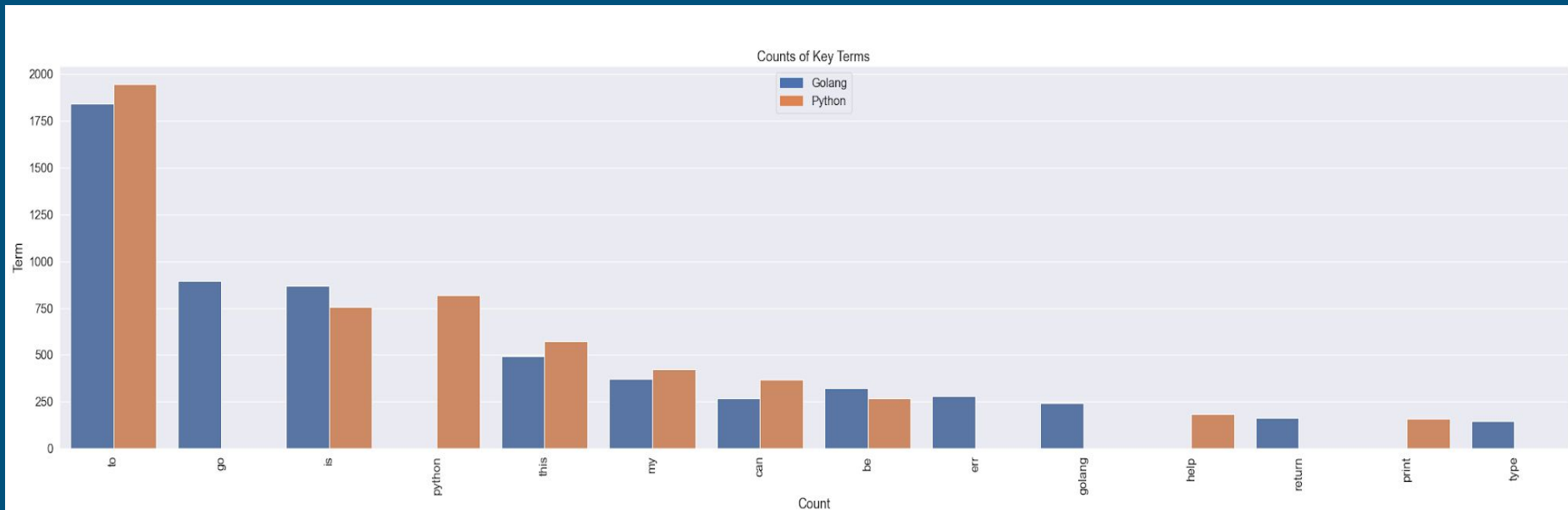






		Training Data	Testing Data
1	Logistic Regression (CVEC)	87%	84%
2	Logistic Regression (TFID)	90%	85%
3	Random Forest (CVEC)	86%	83%
4	Random Forest (TFID)	86%	80%
5	Naive Bayes (CVEC)	87%	84%
6	Naive Bayes (TFID)	86%	84%
	Ensemble (Voting Classifier)	90%	86%

- Six different Classification models run on combined text (body & title).
- Voting classifier used to have each model 'vote' for a classification.



Models used terms like 'python', 'go', 'golang' to classify.

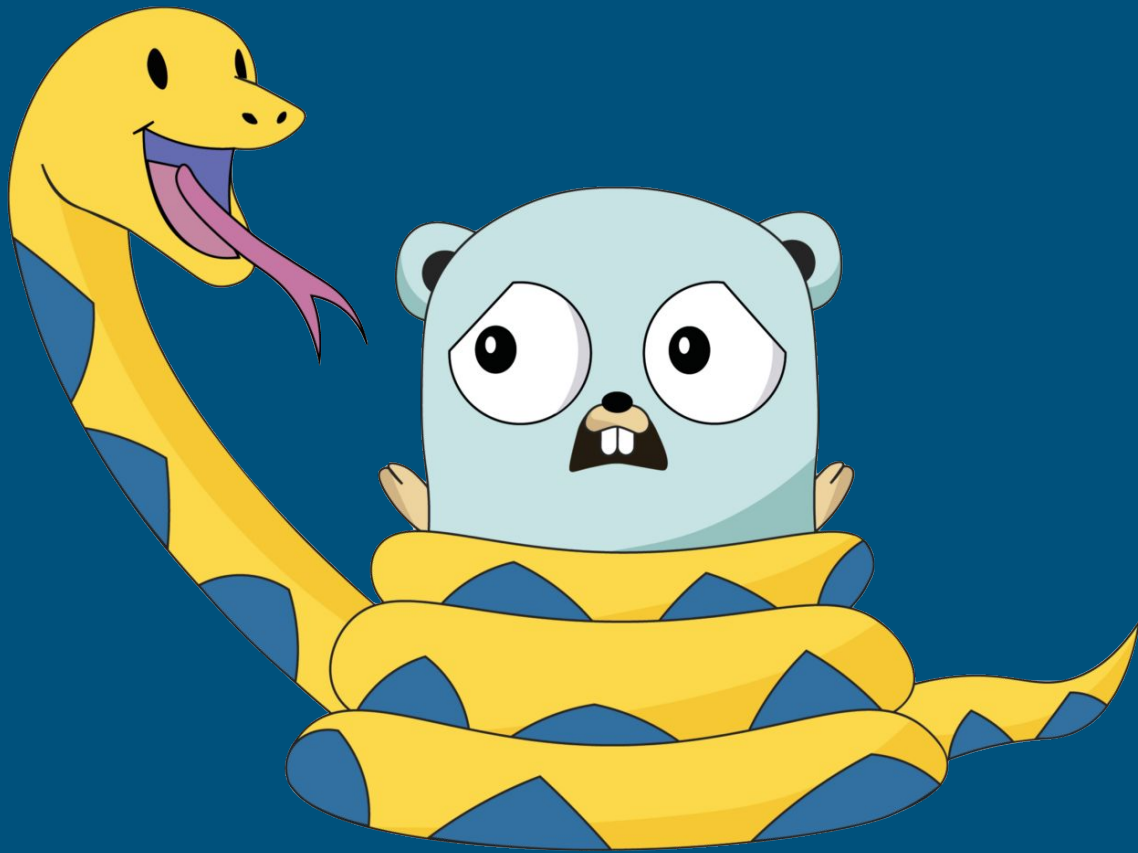
Simplistic but effective.

Conclusions:

- There is no difference between languages wrt sentiment, or word/sentence length.
- Modeling can somewhat accurately distinguish between the two subreddits, but only when including language-specific keywords.

Next steps:

- Remove code from text.
- Get comments and perform same methods.



Questions?