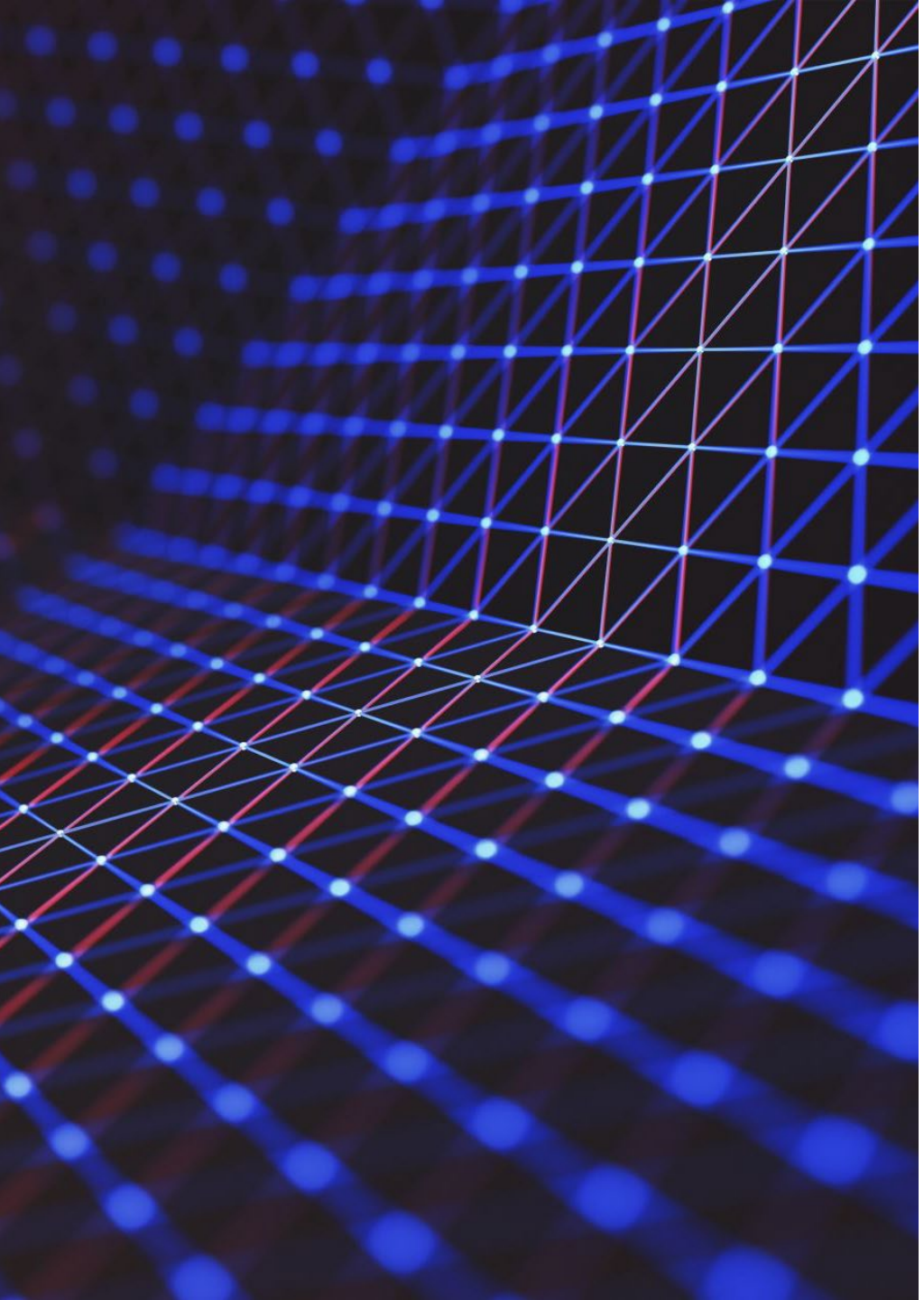
Battle of the Sexes – Blogger Edition

Saint Gau

The Problem

- Incorrect stereotypes: Studies have found that women do not use significantly more words than men, yet these stereotypes persist
- Developing research: Not many studies have been done on gender differences in vocabulary
- NLP for business: Accurate predictors of gender based on text used could be very useful for marketing or other business-related classifications



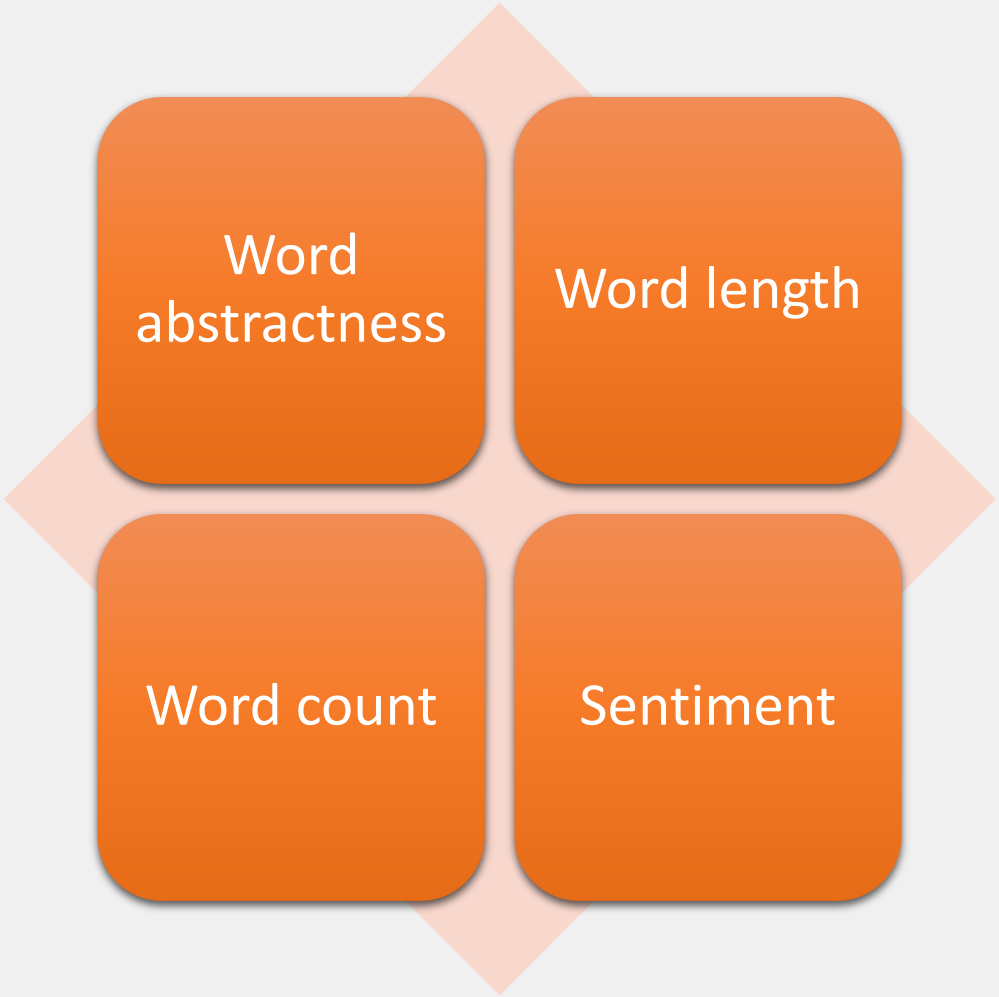
Who might care?

Advertisers

Researchers

General public

How might men and women blog differently?

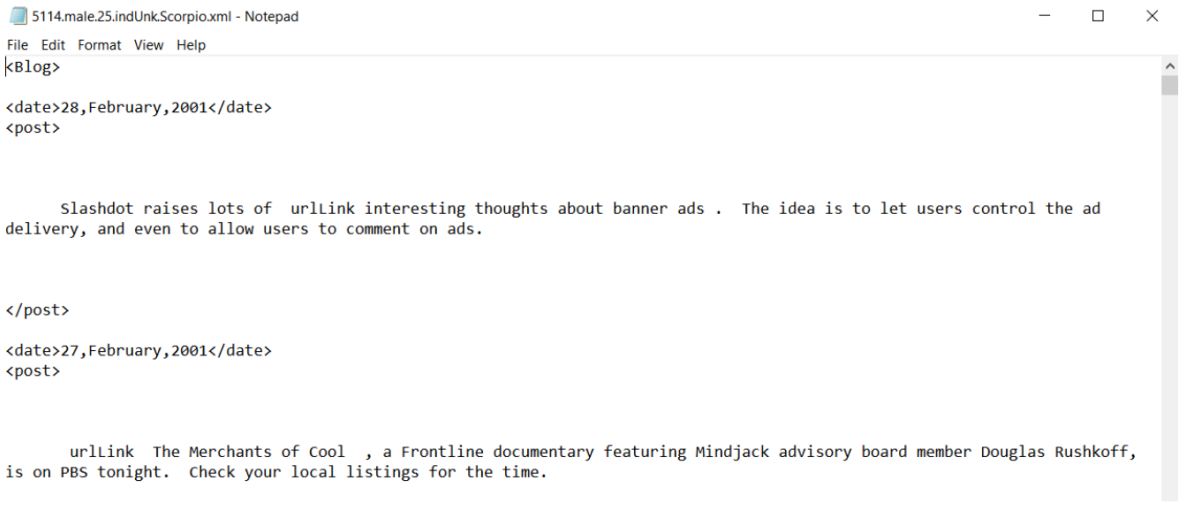


Data Information

Collected August 2004

681,288 blogs written by 19,320 bloggers

Number of fields: 7



	id	gender	age	industry	sign	date	blog
0	1000331	female	37	indUnk	Leo	31,May,2004	Well, everyone got up and going this morning. ...
1	1000331	female	37	indUnk	Leo	29,May,2004	My four-year old never stops talking. She'll ...
2	1000331	female	37	indUnk	Leo	28,May,2004	Actually it's not raining yet, but I bought 15...
3	1000331	female	37	indUnk	Leo	28,May,2004	Ha! Just set up my RSS feed - that is so easy!...
4	1000331	female	37	indUnk	Leo	28,May,2004	Oh, which just reminded me, we were talking ab...

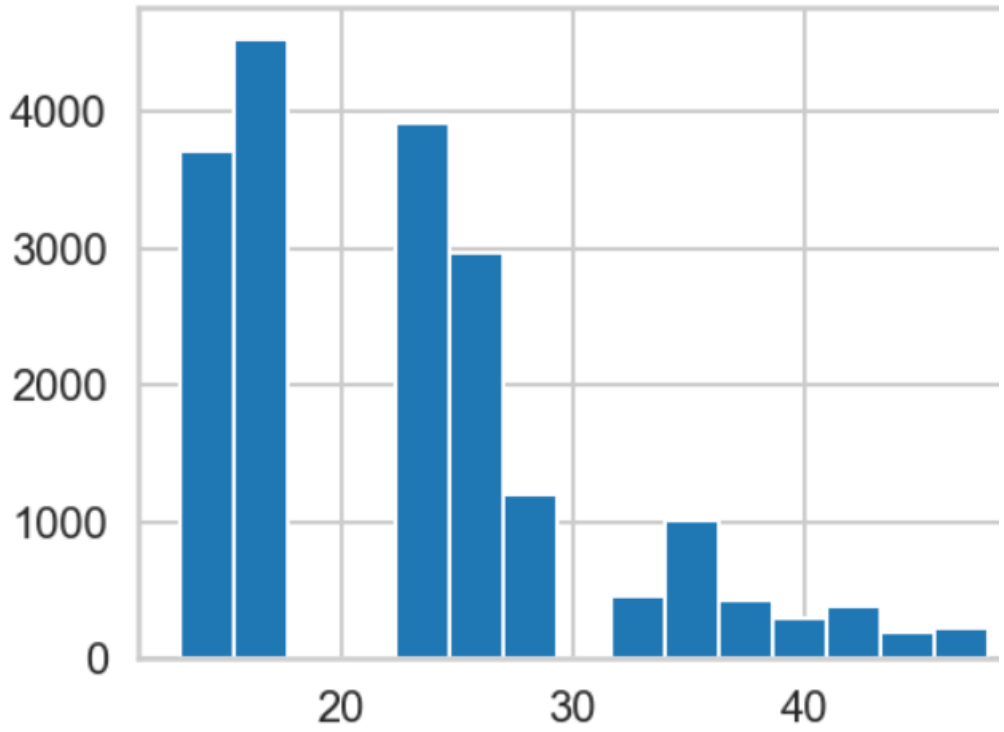
DataFrame Creation

Data Exploration

- Age
- Gender
- Industry
- Date
- Languages
- Sentiment analysis
- Word clouds

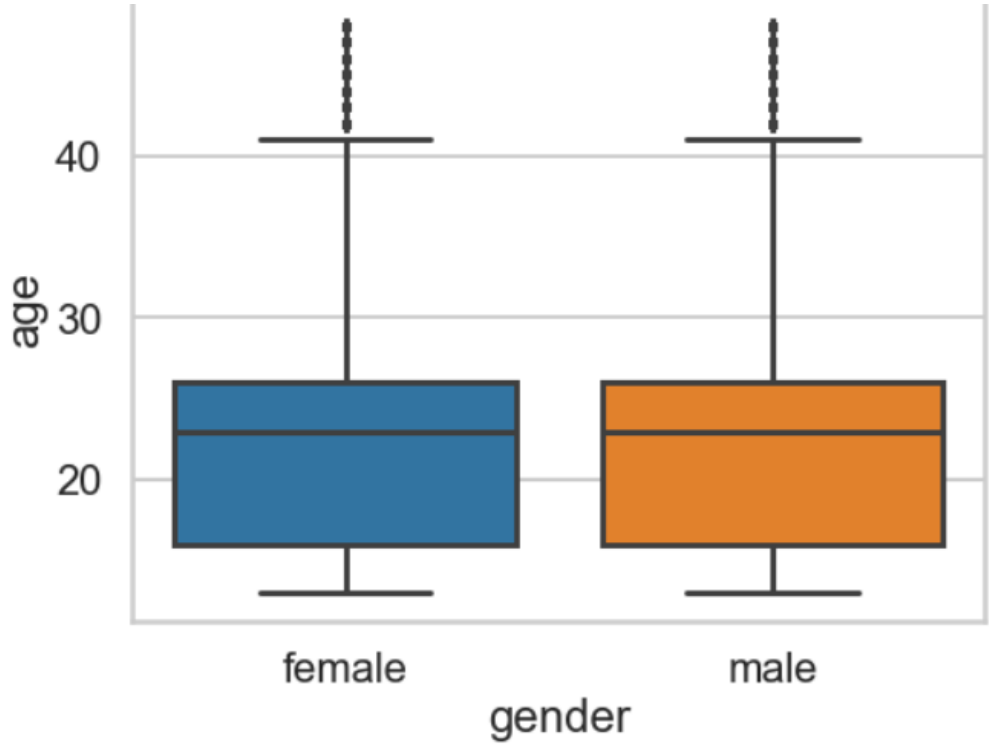
Age

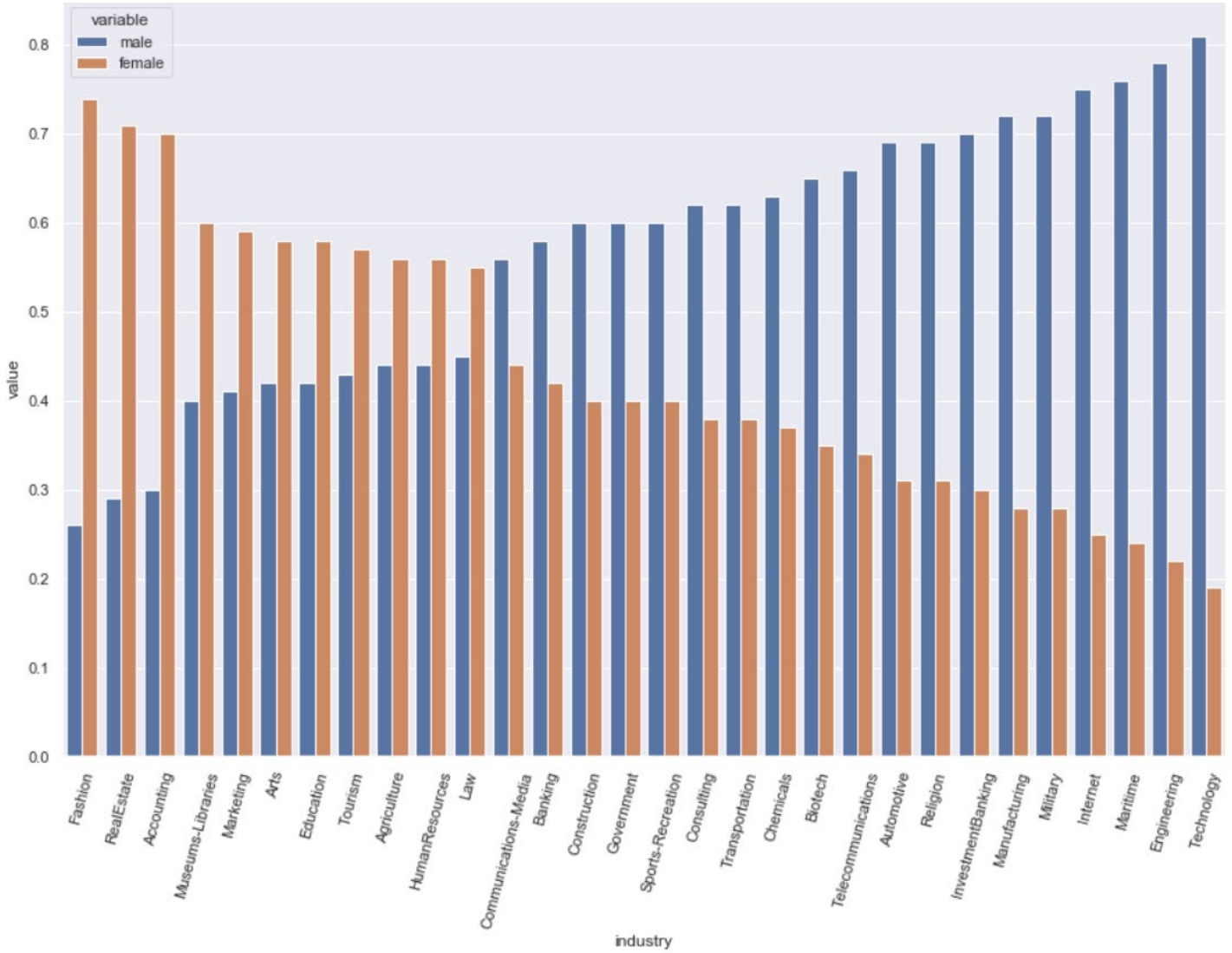
- Gaps in age distribution



Age vs. Gender

- No gender differences in age



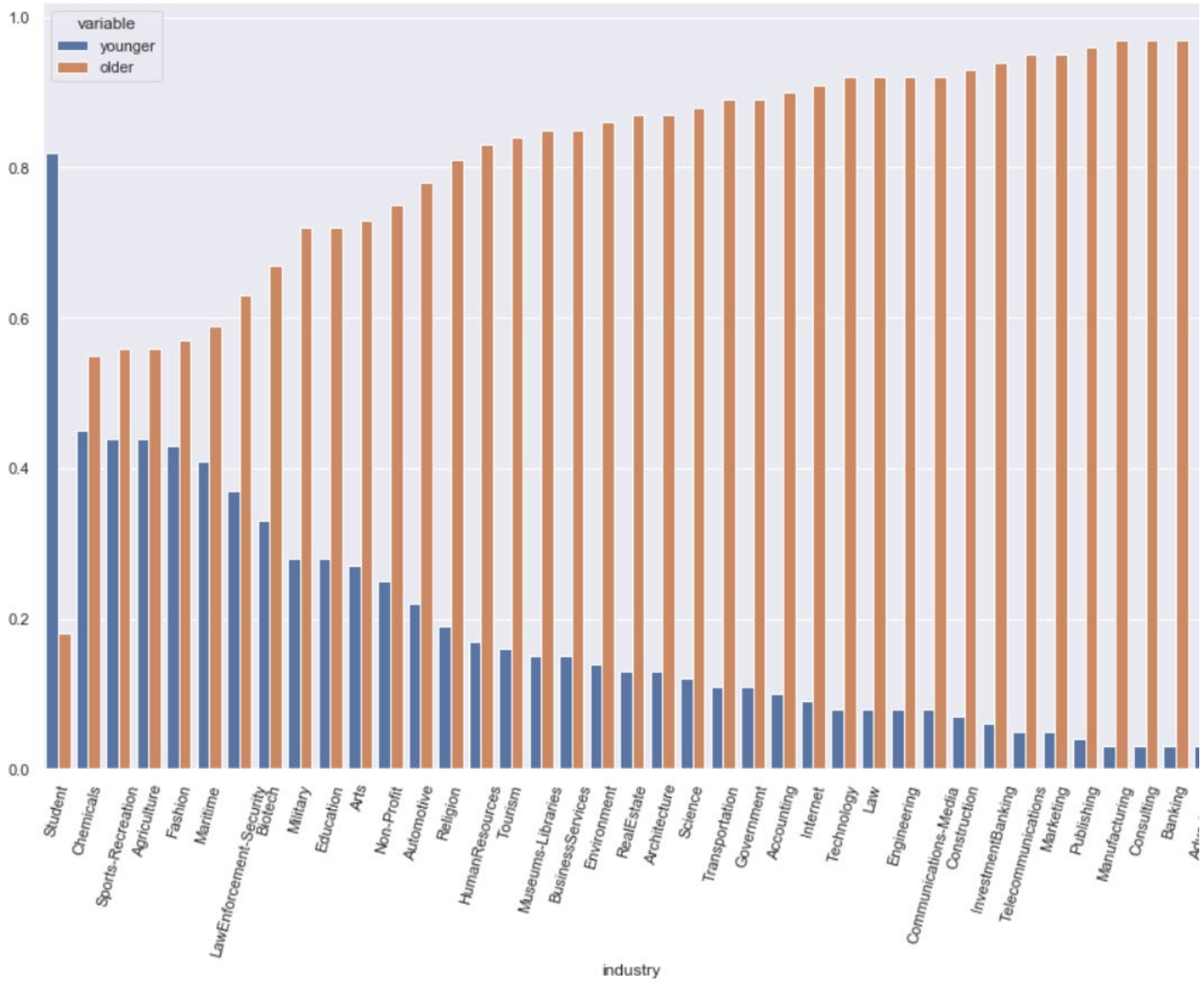


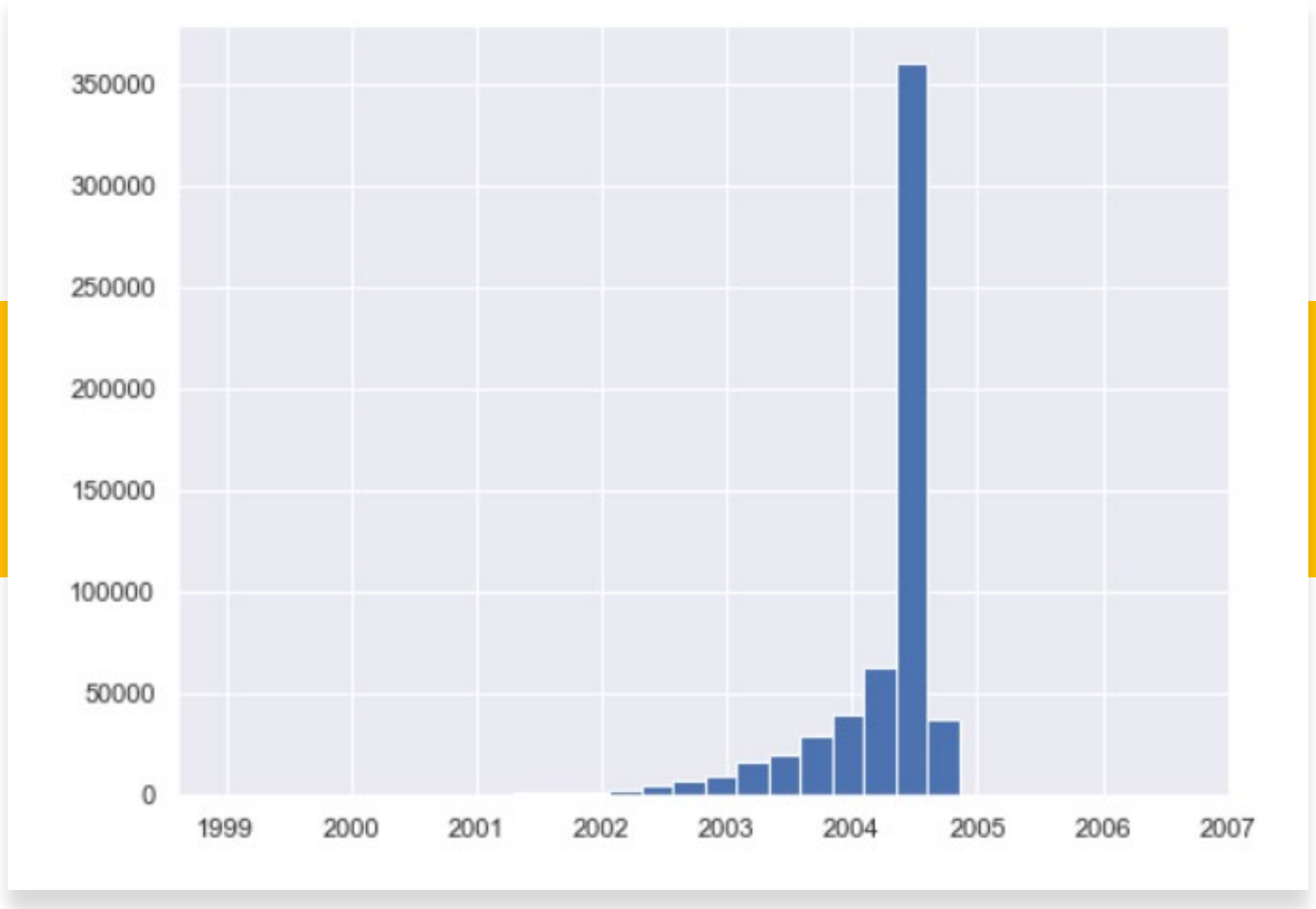
Gender vs. Industry

- Gender differences ranged widely across different industries

Age vs. Industry

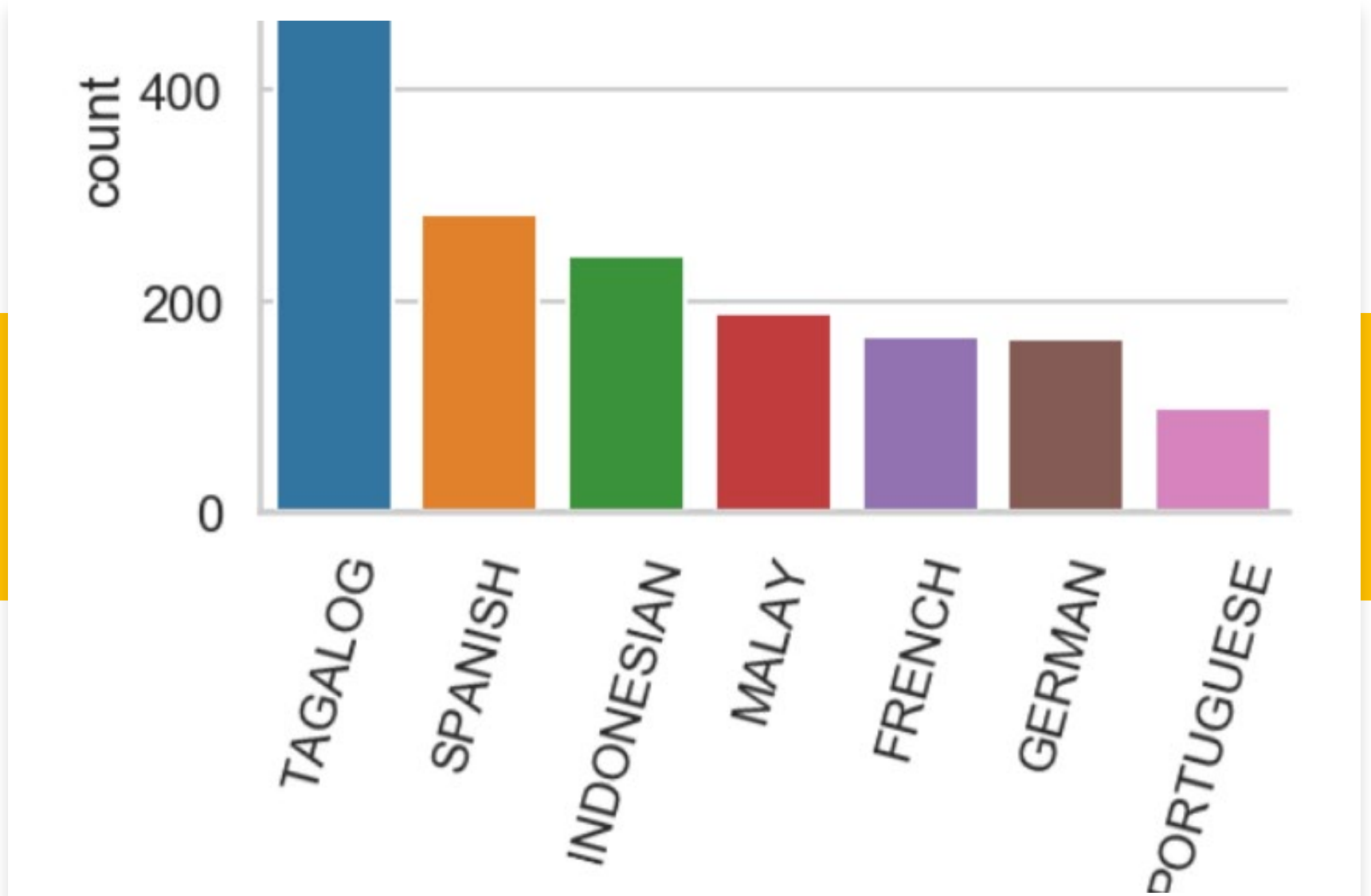
Most industries were made up of mostly older bloggers





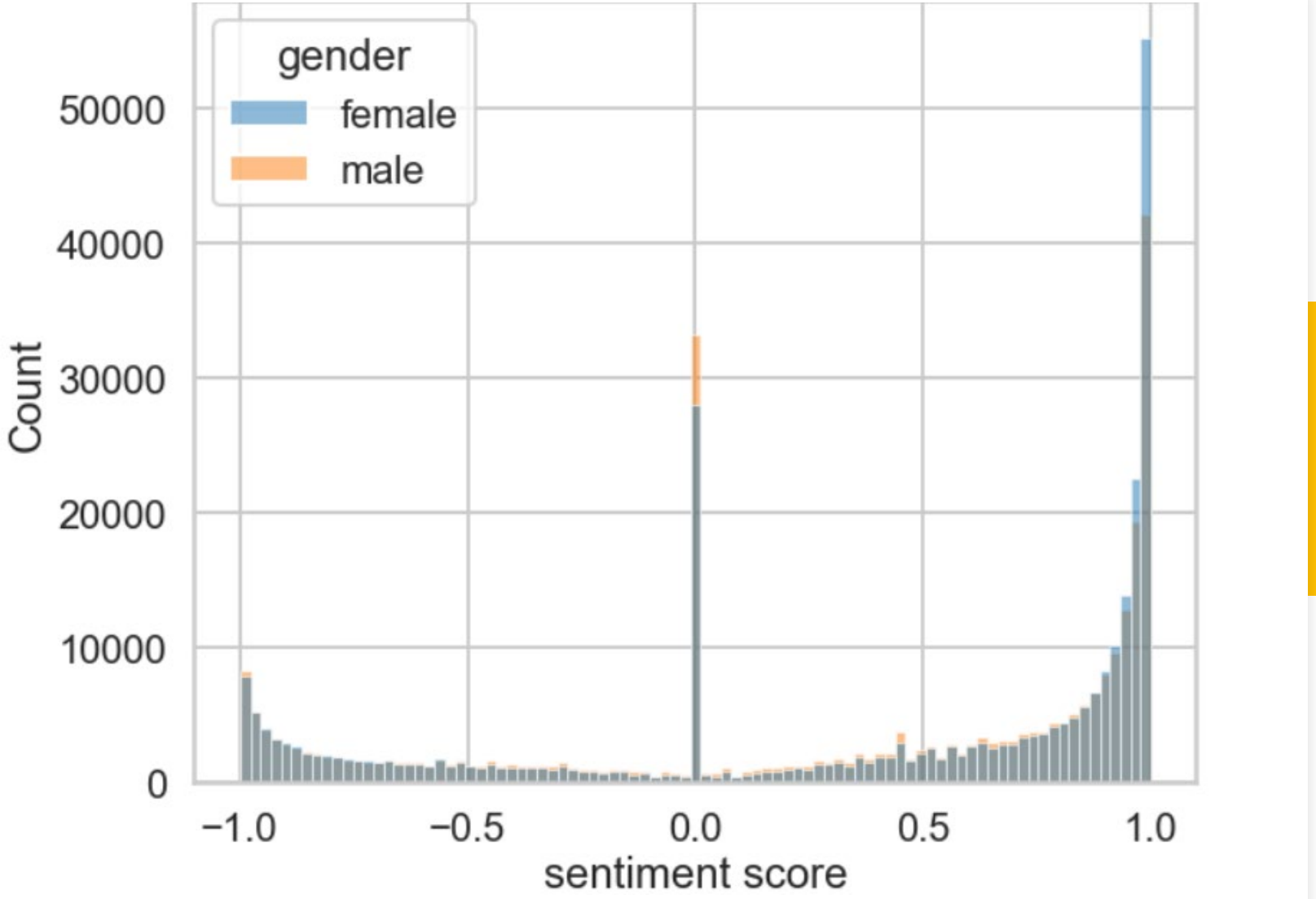
Date

Most blogs were written in 2004



Languages

Most common language of foreign blogs was Tagalog

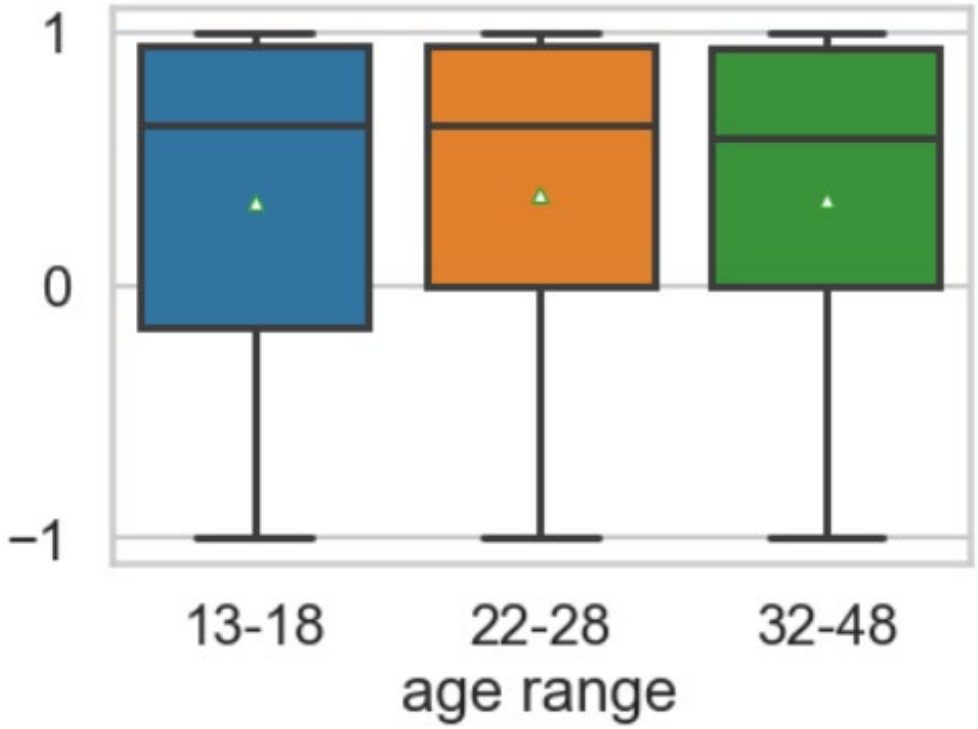


Sentiment Analysis by Gender

Female bloggers were significantly more positive than male bloggers

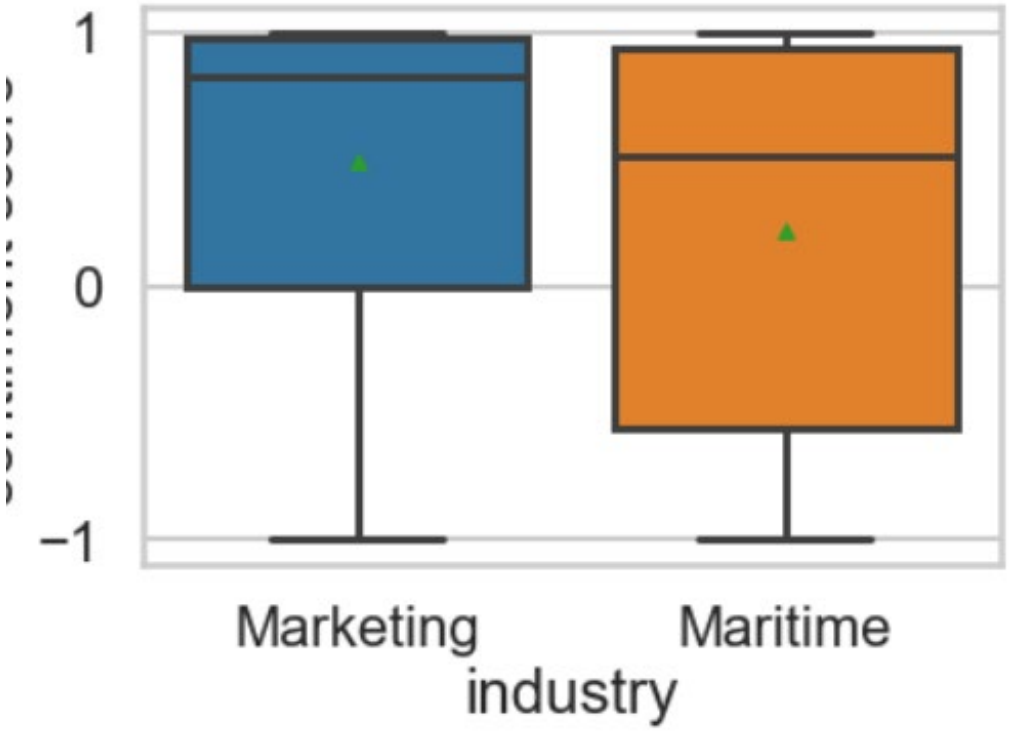
Sentiment Analysis by Age

- 22-28-year-old bloggers were significantly more positive than 13-17-year-old bloggers



Sentiment Analysis by Industry

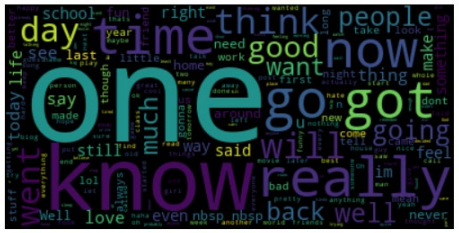
- Marketing bloggers were significantly the most positive, while maritime bloggers were significantly the most negative



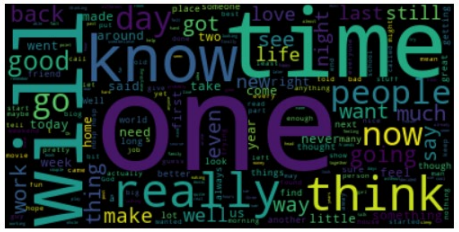
Word Clouds

Word clouds between different groups were mostly similar

Teen word cloud

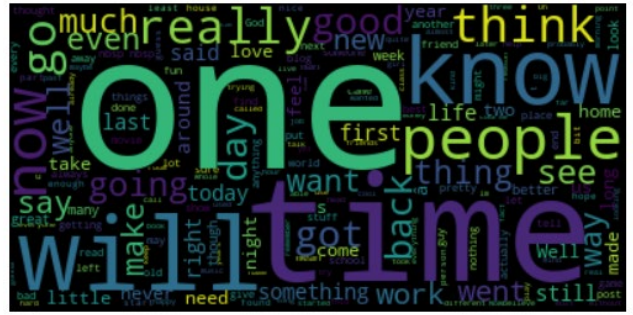


Young word cloud

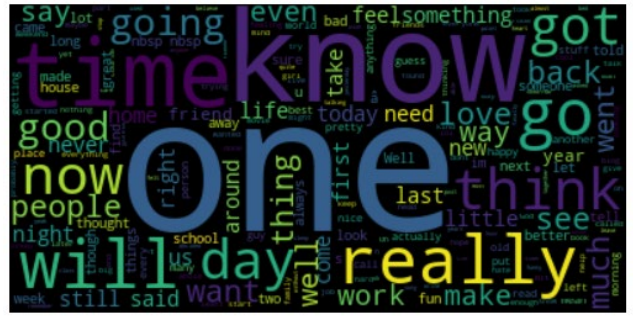


Old word cloud

Male word cloud



Female word cloud



Machine Learning Modeling

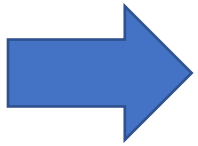
Type: Supervised Learning

Tools: Python's scikit-learn

Dataset too large: only subset used

Modeling Steps

- Pipeline
  - Data Pre-Processing
    - 1. One-Hot Encoding
    - 2. Data splitting into training and test sets (80%-20%)
    - 3. Scaling
  - Cross-Validation (CV) for Hyperparameter tuning
    - 1. 5-fold CV
    - 2. Using scikit-learn's grid search method
    - 3. Evaluation metric: AUC



Performance evaluation using holdout dataset (20% of data used)

Classifier Algorithms Used

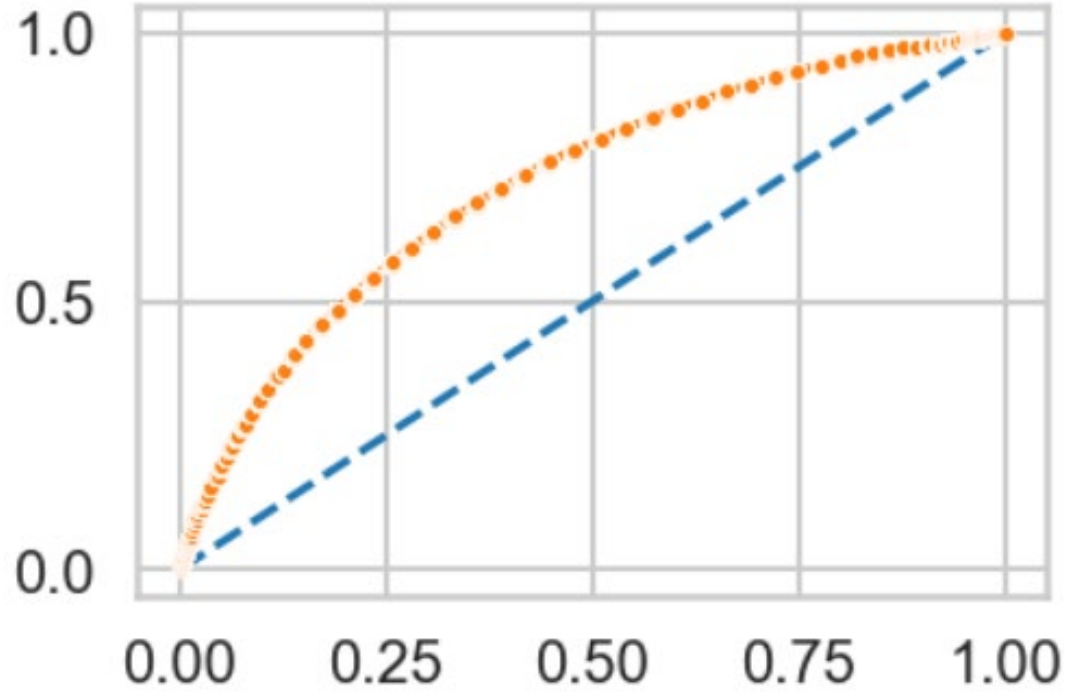
1. Random Forest Classifier

2. XGBoost Classifier

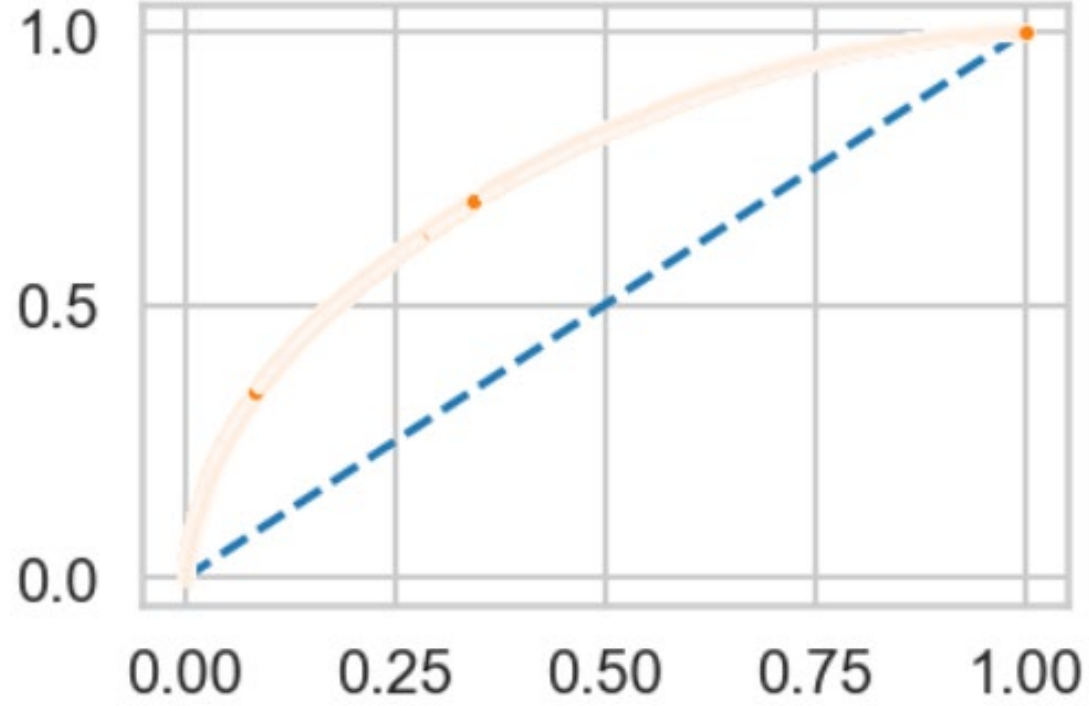
Model Comparisons

XGBoost performed better

AUC: 0.718
Wall time: 37.8 s



AUC: 0.744
Wall time: 3min 17s



Some Details on the Best Model

Tfidf-vectorized

Standard scaling

Using the Model

- Input blog into model
- Use model pipeline on new data and predict gender of blogger

Conclusion

- Men and women do use different words when blogging

Assumptions, Limitations, and Disclaimers

We assume that all bloggers are independent

Blogs were mostly written in 2004, more than 15 years ago

Model would be more accurate if entire dataset were used

More Ideas to Improve the Model in the Future

Diversify information with more features

Extract information from more recent blogs

Include location of bloggers to account for regional differences in language

Thank you!

Saint Gau

Email: transaintgau@gmail.com

https://www.linkedin.com/in/saintgau/

https://github.com/transaint/Professional-Portfolio

Final project report: https://github.com/transaint/Springboard-Projects/blob/master/Springboard%20Projects/Predicting%20a%20Table's%20Tips/Final%20Project%20Report.ipynb