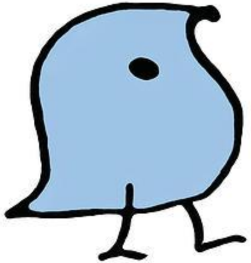
Investigating Gender With NLP Techniques



The REAL Wuglet Research Institute's Pet Shark Happiness Research Group

Introduction/Background

Question: How can NLP tools highlight stereotypes prescribed to gender in texts? Not only in classification, but also in identifying features (ie. passive voice, use of exclamation points, etc...)

Work has already been completed in the computational identification of gender:

- Ethics: Brian Larson, 2017
- Mukherjee and Liu, 2010 have attempted to classify blog entries via POS tagging.
- Montero et al., 2014 extract emotions as features to classify their dataset.

Motivation

Can we use NLP techniques to debunk harmful stereotypes about women’s language use?

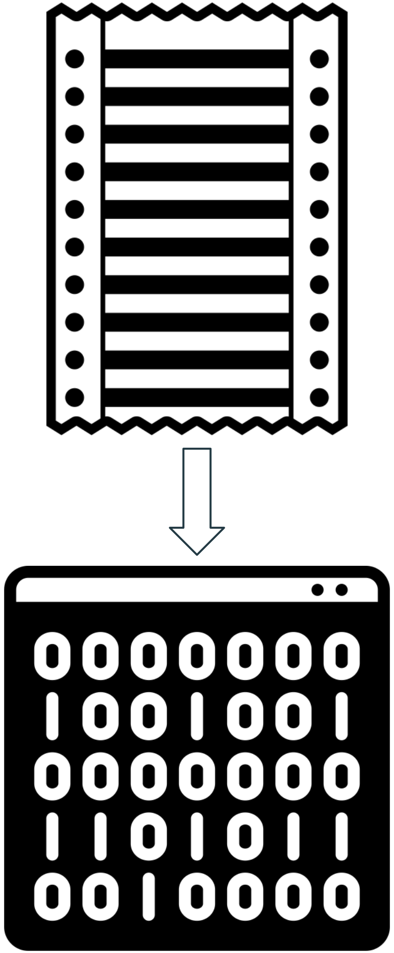
1. How well do programs differentiate between male- and female-authored works? How might we improve upon them?
2. What, if any, features prompt the computer to differentiate between them?

Hypothesis: We can build and train a classification system that will predict whether an author is male or female based on analysis of a text’s sentiment, word choice, and other features to address gender stereotypes.

Broad Strokes: Using baselines of “Stupid” and Naive Bayes, feature engineering and dense vector representations

Our Training Data

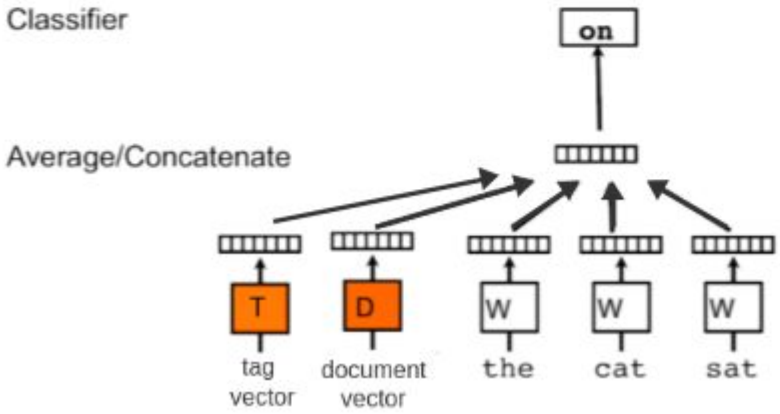
- Blog Dataset
  - Sourced from Mukherjee and Liu, 2010
  - Likely spans multiple years and sites
  - Imported 3096 of the 3227 blog entries
    - 1548 tagged Female
    - 1548 tagged Male
- NYT Dataset
  - Sourced from Soler-Company and Wanner, 2014
  - Subject to copy editing & style guidelines
  - Imported 1672 of the 5230 NYT articles
    - 836 tagged Female
    - 836 tagged Male



Images from https://thenounproject.com

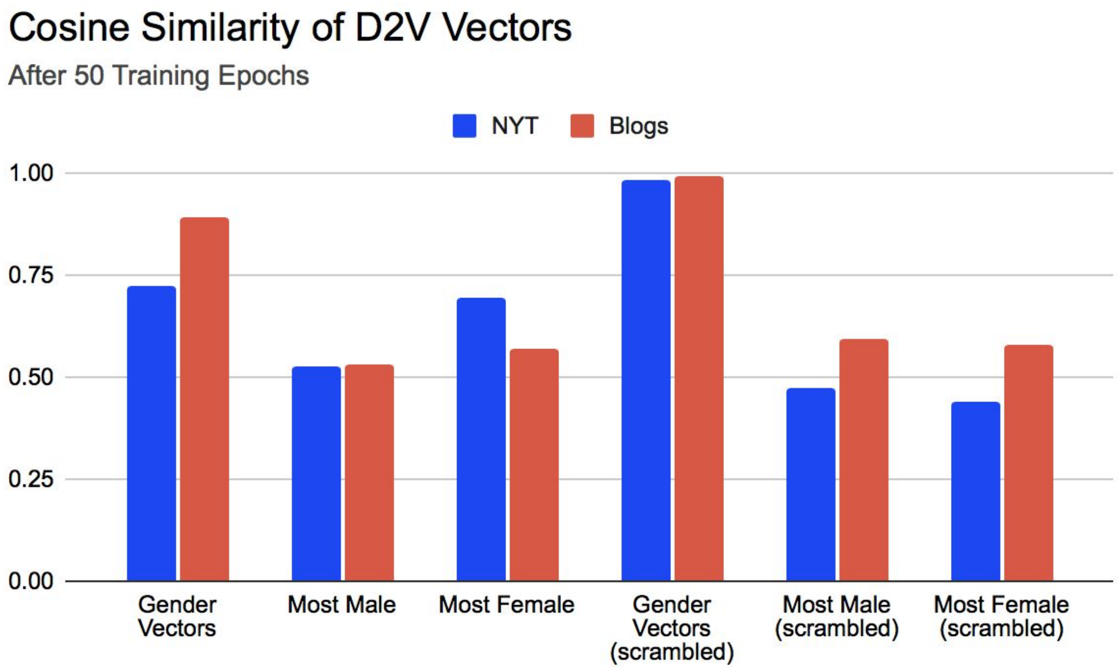
What We’ve Done: Doc2Vec

- Doc2Vec modeling of texts in our corpora.
  - How similar are the texts in our corpora? How different? Is there a significant difference in similarity between documents tagged “M” versus “F”?
  - Parameters:
    - Vector size: 20
    - Training Epochs: 50
    - Alpha: 0.025



Results: Doc2Vec

Measurements of cosine similarity between different vectors of interest vary between trials using “true” tags and trials using tags assigned at random, suggesting that our doc2vec models did in fact pick up on *something*.

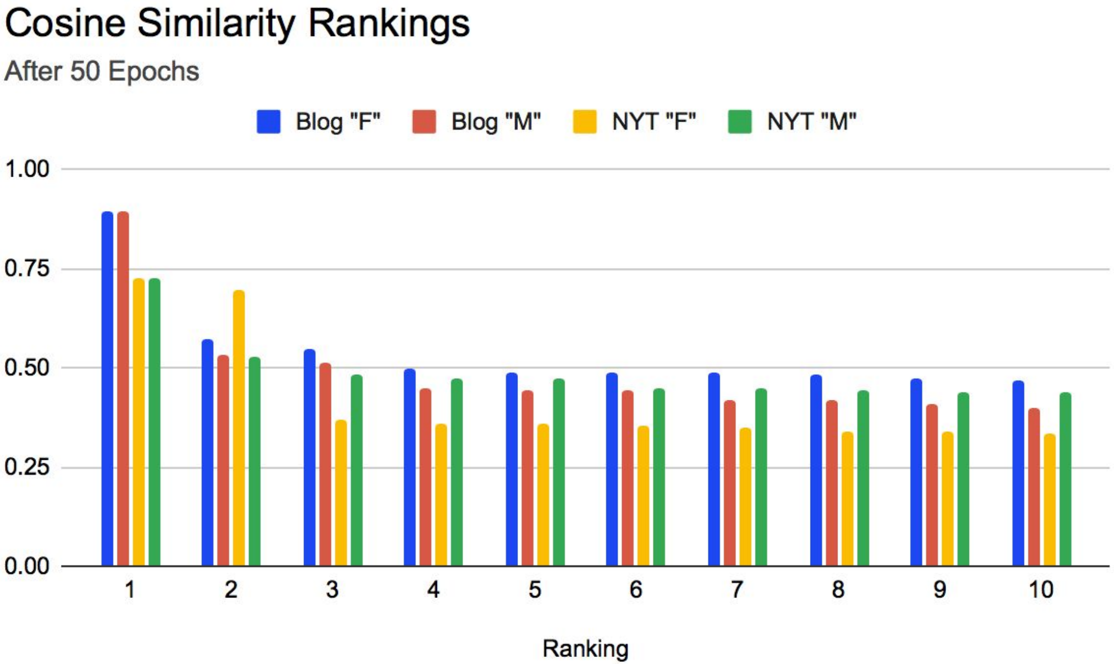


Results: Doc2Vec

So what are the "M" and "F" vectors, and what might we see from measuring cosine similarity of texts against them?

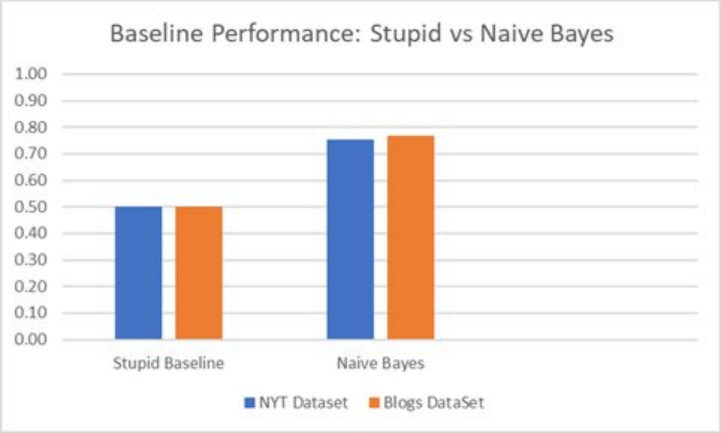
Accuracy: 38.89% for blogs, 94.44% for NYT.

Document 22: high ranking, strange relationship.



What We’ve Done: Baselines

- “Stupid” baseline to compare our different approaches against.
  - Assume all documents are authored by women.
  - Produces 50% accuracy both datasets as our corpora are both split 50/50 by gender.
- Regular Naive Bayes Implementation.
  - Same idea as ham-spam lab.
  - Produces 75% accuracy on NYT data set.
  - Produces 77% accuracy on blog data set.



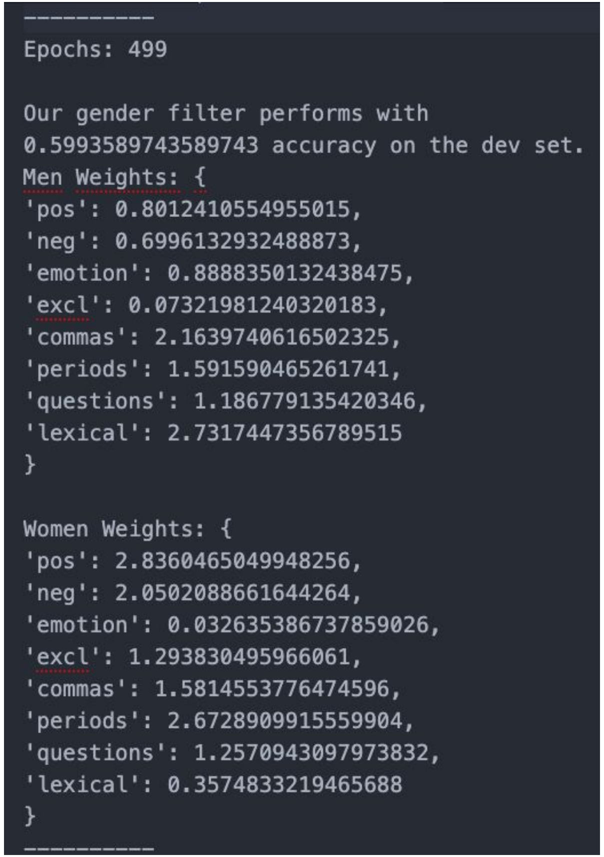
Feature Engineering for our Bayesian Classifier

- Instead of using only terms as features as Naive Bayes does, we honed in on specific aspects of the text
- **In prior research, which features have best predicted author gender?**
  - Soler-Company and Wanner, “How to Use Less Features and Reach Better Performance in Author Gender Identification”
    - Only word- and character-based features
- **How will we determine their weights?**
  - Tuning

Tuning fork image from cleanpng.com

Features We Used

1. **Sentiment analysis**
   a. Women assumed to be more emotional and expressive
   b. Calculated positive/negative connotations of words
2. **Exclamation points (normalized frequency)**
   a. Several studies have found that on average, women use more exclamation points in emails than men do
3. **Other character-based features**
   a. We hypothesized that women and men differ in style, so measured punctuation marks
4. **Lexical diversity**
   a. Experimental!

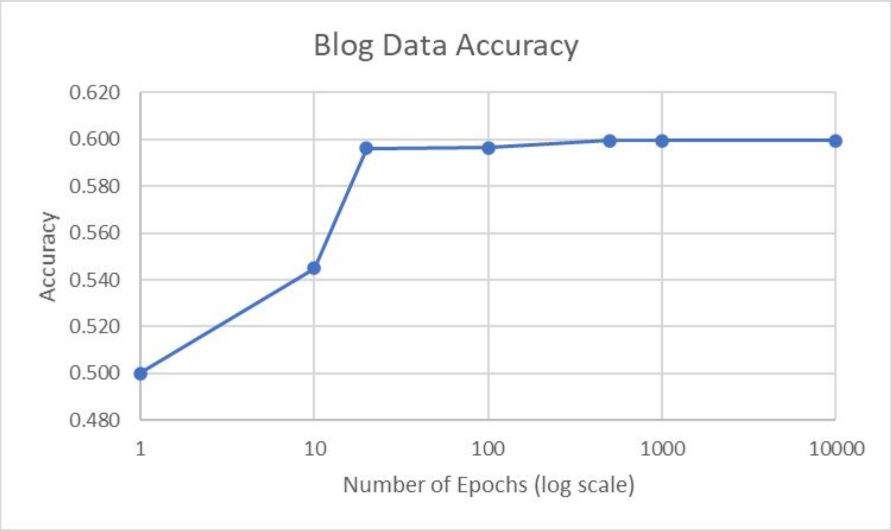


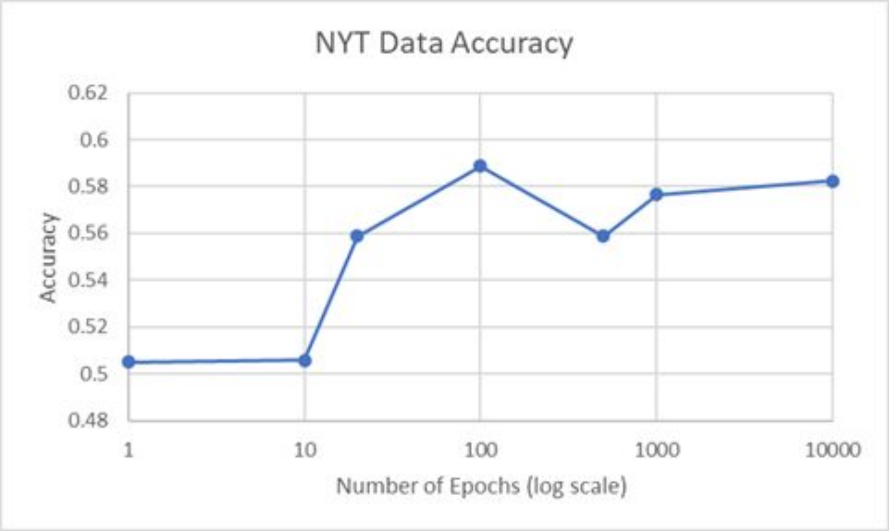
```
----------
Epochs: 499

Our gender filter performs with
0.5993589743589743 accuracy on the dev set.
Men Weights: {
'pos': 0.8012410554955015,
'neg': 0.6996132932488873,
'emotion': 0.8888350132438475,
'excl': 0.07321981240320183,
'commas': 2.1639740616502325,
'periods': 1.591590465261741,
'questions': 1.186779135420346,
'lexical': 2.7317447356789515
}

Women Weights: {
'pos': 2.8360465049948256,
'neg': 2.0502088661644264,
'emotion': 0.032635386737859026,
'excl': 1.293830495966061,
'commas': 1.5814553776474596,
'periods': 2.6728909915559904,
'questions': 1.2570943097973832,
'lexical': 0.3574833219465688
}
----------
```

Tuning: Custom Classifier Accuracy Improvement

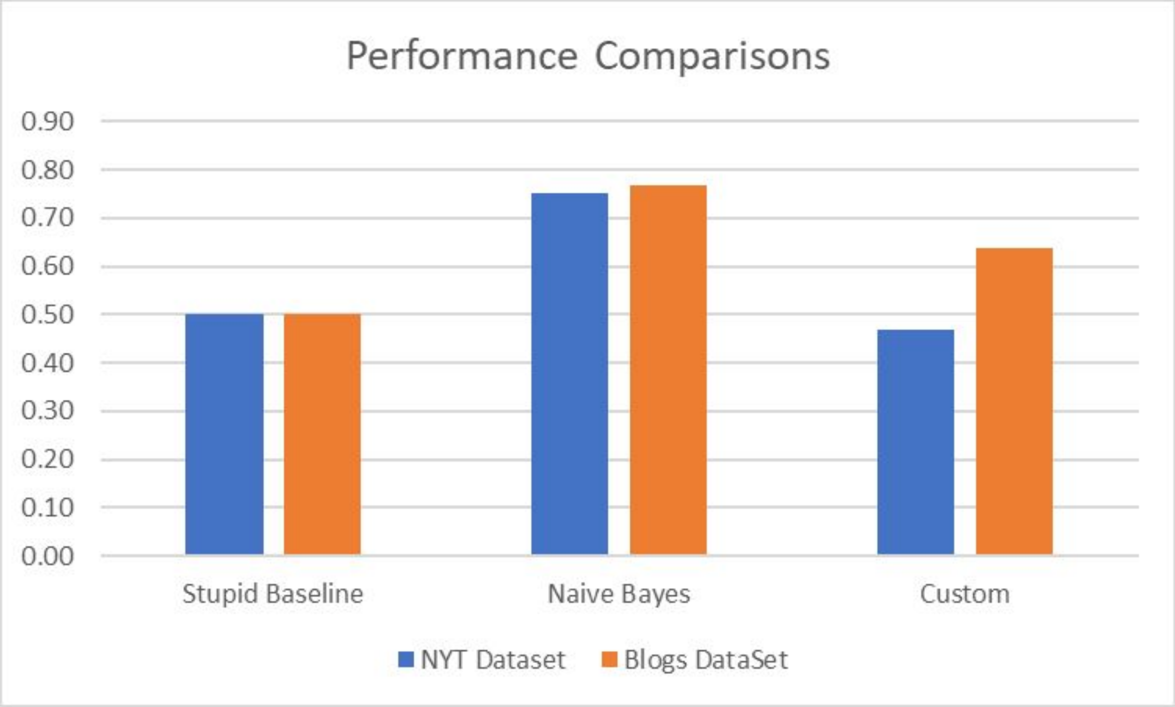




Results: Overall

Ultimately, we could not beat naive bayes. The custom classifier performed with the following accuracy:

- NYT: 63% accuracy
- Blog: 47% accuracy

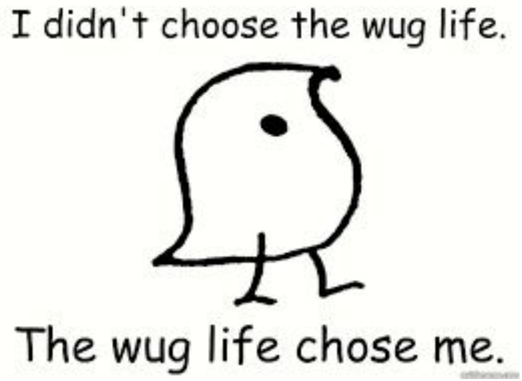


Future Work

- **Compare with other classification systems (NN, log regression)**
- **Include more features:**
  - **More dictionary-based features:** searching for patriotic word usage, curse word usage
  - **Part of speech usage** -- women may use more adjectives
- **Apply to other datasets :**
  - Reddit - in some communities on reddit people identify their gender in their post
- **Genre:**
  - Identify which genres exhibit greatest discrepancy between male-authored and female-authored texts

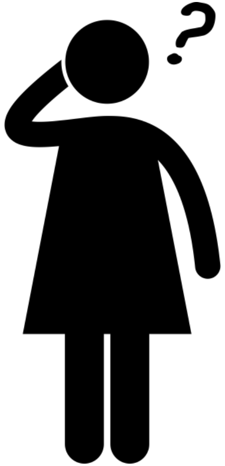
Thank you

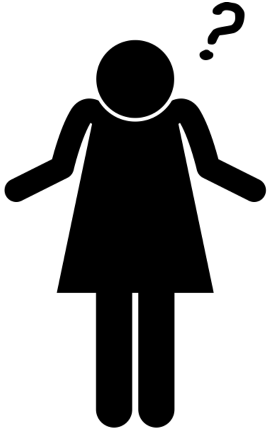
- Professor Caplan
- Dr. Soler-Company
- Viewers Like You
    - For a great semester



https://i.pinimg.com/originals/2e/b4/34/2eb434108298ade6121d090d07c358eb.jpg

Questions?





Images from https://thenounproject.com