Predicting Sexual Orientation Via Facebook Status Updates

Goal and Motivation

34 mins · Stanford, CA

Goal: Predict sexual orientation from Facebook status updates.

Motivation: We want to examine the hypothesis that people with different sexual orientations express themselves differently on social media. Combining our results with our CS 221 Project, which extracted gender features from status updates, we seek to test the stereotype that male homosexuals tend to use more feminine language.

Results

34 mins · Stanford, CA

- For F1 Score, the figures in parentheses indicate F1-scores for homosexuals and heterosexuals respectively.
- SVM Model Parameters Tuning:
  - N-gram range: (1,2), (1,4), (1,5)
  - Min document frequency: 1, 0.95, 0.9
  - Max document frequency: 1, 0.95, 0.9
  - Kernel: Linear, Poly, Rbf

Data

34 mins · Stanford, CA

- We used data from myPersonality.org, with kind permission from Dr. Michal Kosinski (Stanford GSB), which contains 22M Facebook status updates and included demographic details (e.g. gender) of every user in the dataset.
- We derived the sexual orientation labels by looking at the gender of a user’s partner, and comparing it to the user’s gender.
- Word stemming was applied on the status updates.
- Our dataset is skewed in a 9:1 ratio. As such, our test error did not provide a meaningful sense of how our model performed, and we used alternative measures like F-1 score and ROC curves instead.

Features and Models

34 mins · Stanford, CA

- N-grams (tuned across a range of hyper-parameters)
- Counts of periods, exclamation marks, smileys and capital letters.

Learning Algorithms:

- Support Vector Machine
- Multinomial Naïve Bayes
- Logistic Regression
- Random Forest

Analysis and Future Work

34 mins · Stanford, CA

- ROC scores for both females and males were above 60%, which told us that there were distinctions in how homosexuals expressed themselves on social media, even if the distinction was not great enough to consistently predict one’s sexual orientation.
- Mentions of another partner of the opposite gender (e.g. when males mention ‘wife’) are strong indicators that a person is homosexual.
- Our model showed that a male homosexual was 4 times more likely to use the word “gay.” In fact, a male who mentions “gay” in a status update has a 1 in 4 chance of being homosexual.
- Future Work:
  - Exploration of other methods of feature extraction (e.g. Word2Vec), and more nuanced feature engineering. We can also use neural networks to automatically learn features in the data.