Predicting Subreddits: ASD vs. OCD

GA DSI Project 3
Julia Kelman

Identifying Potential Customers

Problem Statement

We created a **product** meant to help individuals with **Autism Spectrum Disorder (ASD)** and want to **market it online**. We need to **identify potential customers** based on their online content.

Other disorders like **Obsessive Compulsive Disorder (OCD)** share many symptoms with autism. As a result, online resources are often **geared towards both of ASD and OCD**.

We need to be able to differentiate between people with ASD and OCD based on what they post on online platforms.

We plan to solve this problem by using submissions on an **Autism reddit page** and an **OCD reddit page** to build a **classification model** able to **classify a user as having either ASD or OCD** based on their post with the **highest level of accuracy** possible.

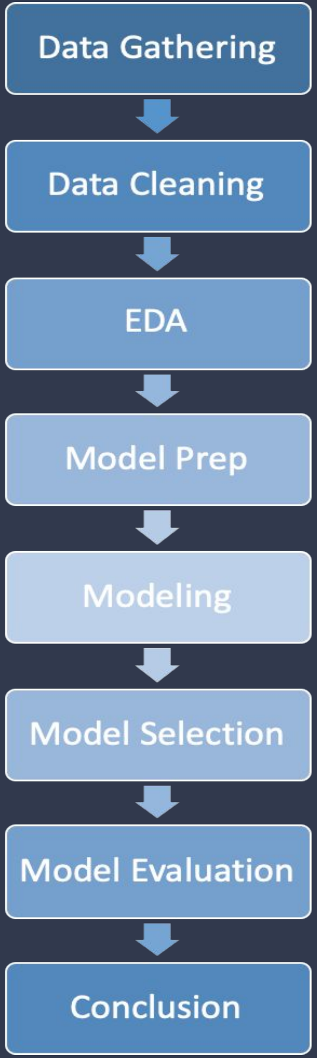
Data:
The origin

From: Reddit

Posts submitted by individuals between Nov. 2019 and March 2020

1653 ASD Submissions
2209 OCD Submissions
8 Variables

Workflow



Data Gathering

Data Cleaning

EDA

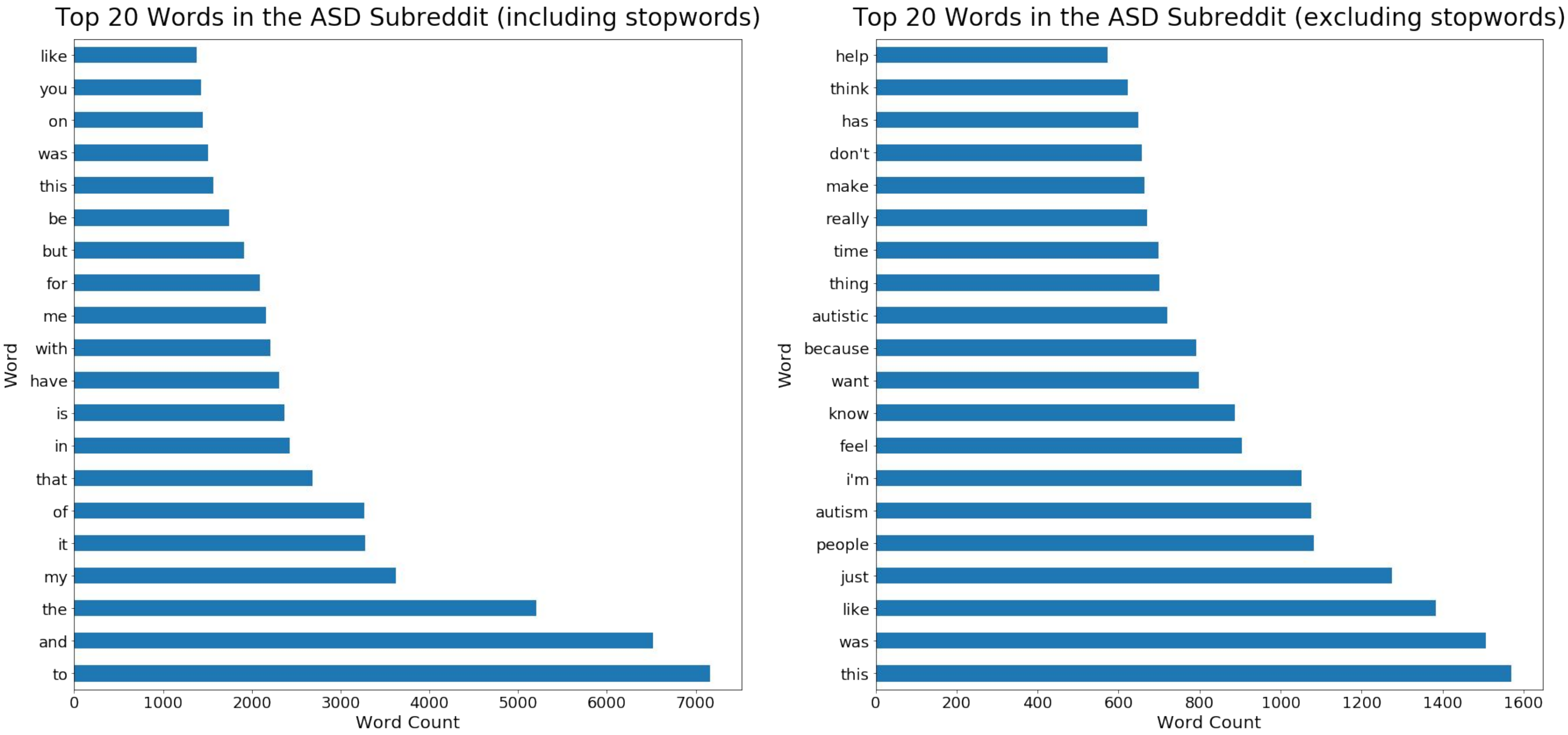
Model Prep

Modeling

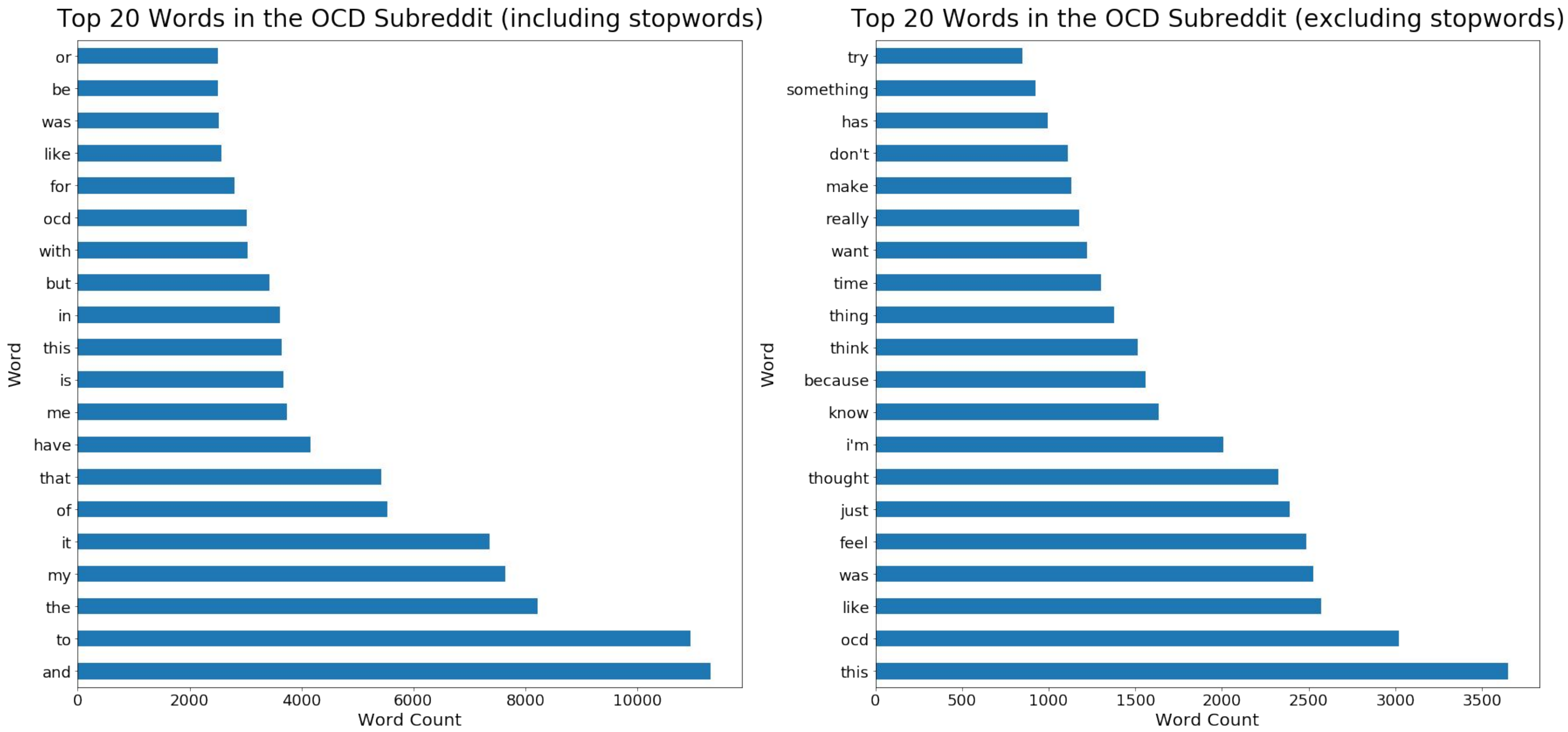
Model Selection

Model Evaluation

Conclusion



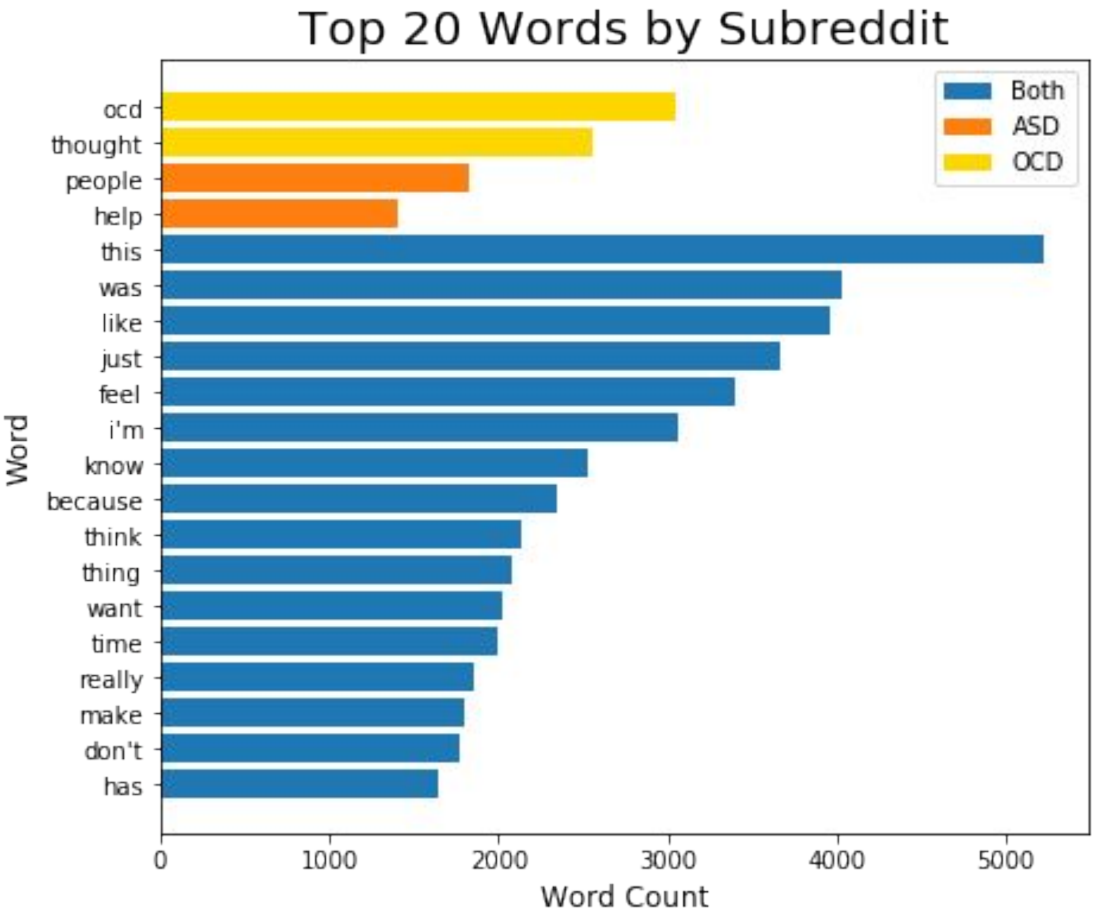
20 MOST FREQUENT WORDS IN THE ASD SUBREDDIT



20 MOST FREQUENT WORDS IN THE OCD SUBREDDIT

20 Most Frequent Words Overall

- Most words in both subreddits
- Need for custom stopwords
- Subreddit specific words emerge



Modeling

X: Cleaned Text
(Title + Self Text)

Y: Subreddit
(ASD or OCD)

9 Models:

- Baseline
- Logistic Regression + Cvec
- Logistic Regression + TFIDF
- kNN + Cvec
- kNN + TFIDF
- Multinomial Naive Bayes + Cvec
- Gaussian Naive Bayes + TFIDF
- SVM + Cvec
- SVM + TFIDF

Model Selection

Based on Accuracy

	Baseline	Logistic Regression + Cvec	Logistic Regression + TFIDF	Knn + Cvec	kNN + TFIDF
Accuracy	Train: 0.5718 Test: 0.5725	Train: 0.9917 Test: 0.9213	Train: 0.9889 Test: 0.9337	Train: 0.7935 Test: 0.7298	Train: 0.6057 Test: 0.5818

	Multinomial Naive Bayes + Cvec	Gaussian Naive Bayes+ TFIDF	SVM + Cvec	SVM + TFIDF
Accuracy	Train: 0.9437 Test: 0.9141	Train: 0.9579 Test: 0.9037	Train: 0.9772 Test: 0.8913	Train: 0.9945 Test: 0.8996

- Predicting with 93% Accuracy

Model Selection

Based on Accuracy

	Baseline	Logistic Regression + Cvec	Logistic Regression + TFIDF	Knn + Cvec	kNN + TFIDF
Accuracy	Train: 0.5718 Test: 0.5725	Train: 0.9917 Test: 0.9213	Train: 0.9889 Test: 0.9337	Train: 0.7935 Test: 0.7298	Train: 0.6057 Test: 0.5818

	Multinomial Naive Bayes + Cvec	Gaussian Naive Bayes+ TFIDF	SVM + Cvec	SVM + TFIDF
Accuracy	Train: 0.9437 Test: 0.9141	Train: 0.9579 Test: 0.9037	Train: 0.9772 Test: 0.8913	Train: 0.9945 Test: 0.8996

Model Evaluation: Confusion Matrix

	Predicted OCD	Predicted ASD
Actually OCD	516	37
Actually ASD	27	386

Correctly classifying:

- **516 / 553** OCD posts
  True Negative Rate: 93%
- **386 / 413** ASD posts
  True Positive Rate: 93%

Marketing:

- **3.8%** of cases: spending advertising money on wrong users
- **2.8%** of cases: missing out on potential customers

Model Evaluation: Understanding Misclassification

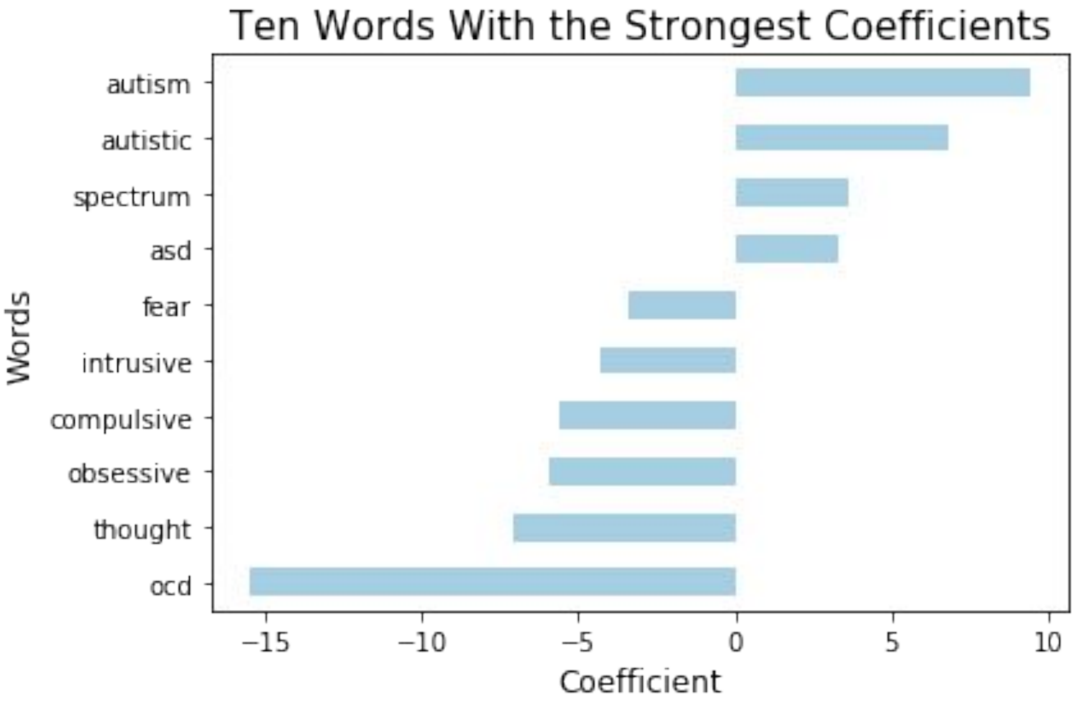
“Anyone buy Sheryl Paul’s books? Have they **helped**?”

Predicted: **ASD**

Actually: **OCD**

Model Evaluation: The Words

- 1 unit increase in use of the word "**Autism**" -> **11,938 times as likely** to be posted in the ASD subreddit
- 1 unit increase in use of the word "OCD" -> 99.9% less likely to be posted in the ASD subreddit



Conclusion

Logistic Regression model with **TFIDF Vectorizer** and **custom stopwords** gives us the **highest predictive power** with **93% accuracy.**

In **3.8%** of cases we're spending advertising money on wrong users. In **2.8%** of cases we're missing out on potential customers

References

1. https://www.reddit.com/r/autism/
2. https://www.reddit.com/r/OCD/
3. https://www.webmd.com/brain/autism/autism-similar-conditions