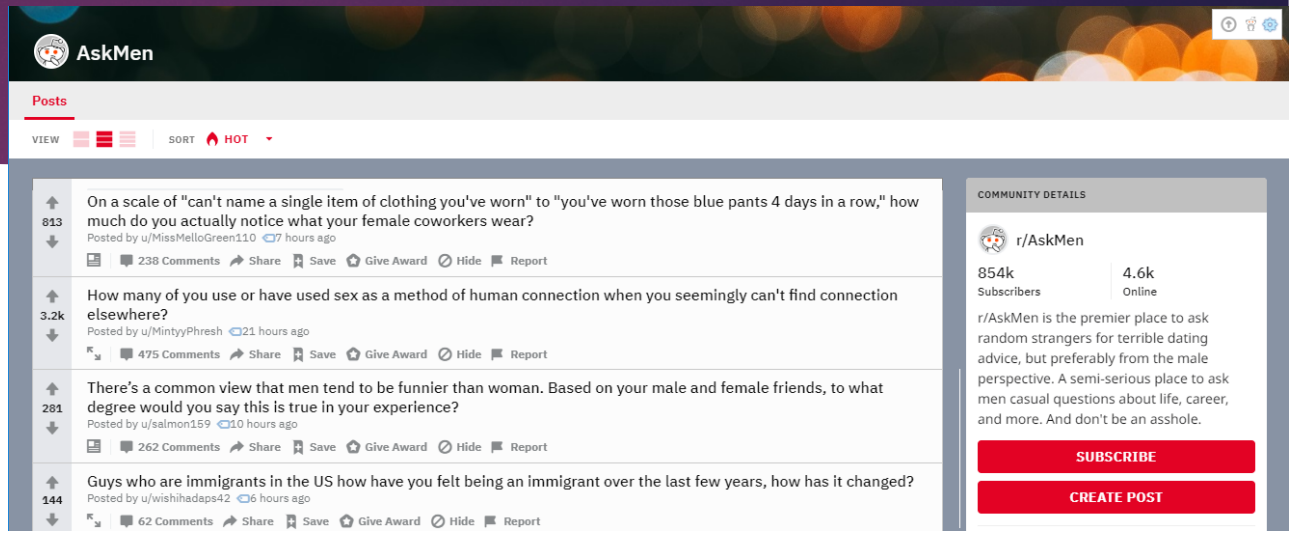
Eamon Fleming – Data Science Immersive – Project 3

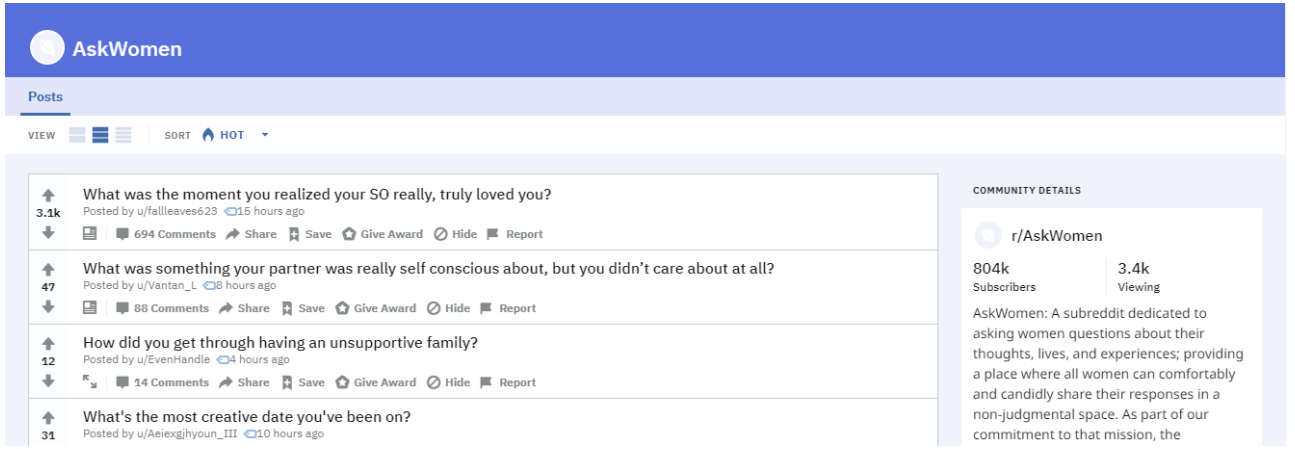
Asking Men and Women

CLASSIFICATION OF SUBREDDIT COMMENTS WITH NATURAL LANGUAGE PROCESSING

A TALE OF TWO SUBREDDITS

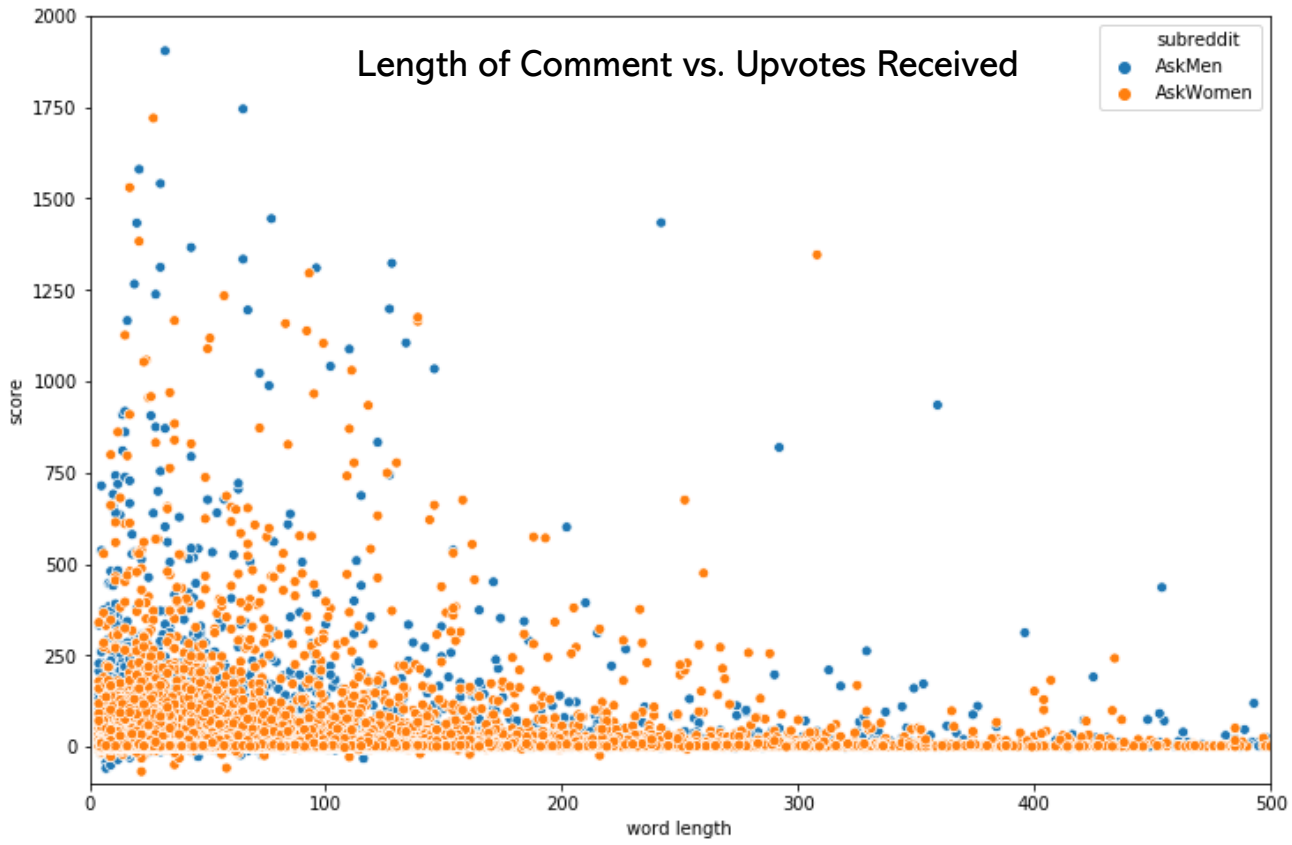
- **r/AskMen** and **r/AskWomen**: two subreddits designed so that redditors can ask questions that they want to direct specifically to redditors of the male or female gender.
- First-tier comment replies in threads should be overwhelmingly women in r/AskWomen, and men in r/AskMen, respectively.
- If we can collect first-tier comments only, from each subreddit, we can use the subreddit label as a proxy for author gender.
- This allows us to build a model to try and predict the gender of Reddit comments.





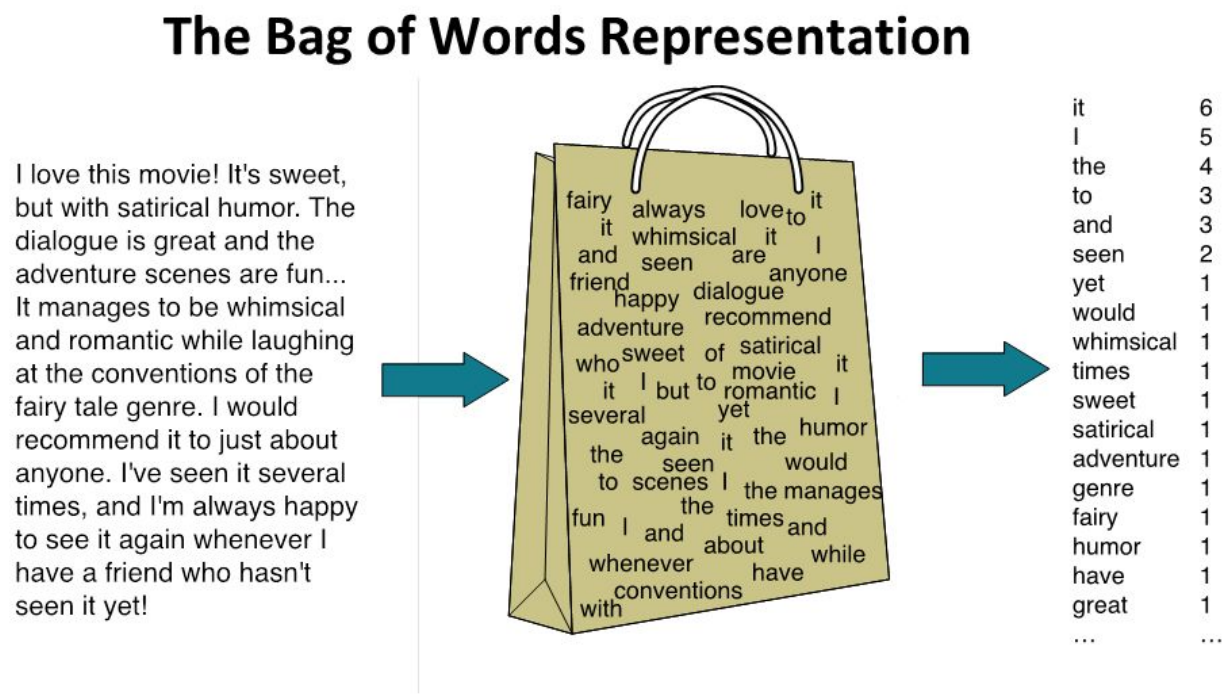
DATA CLEANİNG AND PROCESSİNG

- **Pushshift** is an open-source alternative to Reddit's API that allows us to pull comment data going back in time, and returns data in a more convenient format.
- Pushshift formats parent ids for each comment in a way that lets us **filter for first-tier comments only**.
- **Collected** 1000 first-tier comments from each subreddit at 12-day intervals, going back two years
- **Cleaned away** duplicates, moderator boilerplate, and [deleted] or /removed/ type comments.
- Engineered an accurate **word length** feature using Regex. Excluded comments with <4 words.
- Final clean dataset contains **~35000 comments** from each subreddit.



NATURAL LANGUAGE PROCESSİNG

- **Bag of Words:** we use a **vectorizer** to split comments up into words and convert each comment into a vector of word frequencies (relative positions of words are ignored).
- **CountVectorizer** creates pure frequency vectors. **TfidfVectorizer** normalizes frequencies, down-weighting the influence of common words, and up-weighting rare words.
- We can exclude words that are unlikely to be predictive because they appear too often, or they appear too rarely.
- In addition to individual words, we can look at **n-grams** (pairs or groups of words). This gives us *some* characterization of how words are positioned relative to each other.



MODELING WITH LOGISTIC REGRESSION

- **Logistic Regression** is a classic technique in large-scale classification problems, that has the advantage of being highly **interpretable.**
- We can look at which words or n-grams the model associates most to a gender. We can also look at which comments are easiest for it to gender-identify
- We use **GridSearch** to look at different combinations of vectorizer options and Logistic Regression options, and find the best performing fit.
- On new data, our model predicts the correct gender **70.5%** of the time (baseline accuracy was **53.9%**)! This beats Naïve Bayes and Random Forest models.
- The **confusion matrix** (right) shows that male comments are harder for the model to identify than female comments.

	precision	recall	f1-score	support
AskMen	0.70	0.63	0.66	7755
AskWomen	0.71	0.77	0.74	9084
avg / total	0.70	0.71	0.70	16839

	Predicted AskMen	Predicted AskWomen
Actual AskMen	4879	2876
Actual AskWomen	2089	6995

R/ASKWOMEN DEAD GİVEAWAYS

- "My boyfriend! I love his natural body scent."
- "Cleanse, tone, moisturize, eye primer, shadow, liner, powder foundation, highlight, blush, setting spray, eyebrows, lipstick. Sometimes mascara, although I mostly hate it."
- "My boyfriend wears vintage black and it's so freaking sexy lol. I rarely wear perfume, I prefer natural smells or sweet smells if I do though, pink sugar was my jam in high school lol. I mostly just use lavender and fruity smelling soaps and it's not overpowering but I still smell delicious lol"
- "My boobs grew and hurt in ways I'd never known before and then I started getting really bad cramps but no period. So I took the test while at work and it was positive. A tiny part of me was very excited but I was mostly sad and scared. I have severe endometriosis and was told if I ever got pregnant that the baby would die. So I spoke with my husband and the doctor and we agreed (very reluctantly on our part) on having a medical abortion. That was last July. I switched doctors early this year and was told there's no reason I wouldn't be able to have children. I'd just have to be on bed rest and monitored very closely. I could have had a baby by now and it makes me sad every single day."
- "Cats Zumba My husband Reselling designer clothing"

R/ASKWOMEN REPEATS MATTER

"I'm a graphic designer. For my senior show I had to showcase my work in a gallery. So I went to a few print shops and talked to managers about their quality and printing process. One print shop I went with my boyfriend after lunch, and we walked in and my boyfriend stood behind me while I talked to the manager. I asked a few questions but the manager would direct his answer to my boyfriend. And even asked my boyfriend if he wanted a quote. My boyfriend said "you need to talk to her, it's her project and her money" and my boyfriend went and sat down behind me. I continued asking about color quality, and asked what happens if I'm not happy with the outcome, and the manager laughed and said "well it's not like picking wallpaper for the kitchen sweetheart" and looked at my boyfriend for approval on his joke. He didn't get approval and we left."

R/ASKWOMEN LINGUISTIC PATTERNS

- Transition cadence to further explanation:
  - “and then”, “although”, “also”, “so”, “because”
- Personal touch:
  - “ask me”, “how are you”, “please”, “feel”
- Short and sweet casual language:
  - “super”, “ugh”, “comfy”, “cute”, “lol”, “yum”
- Miscellaneous expressions:
  - “uncomplicated”, “killing my”, “lot more”, “obnoxious”, “would never”, “if someone”, “struggling”, ”my most”, “absolutely”, “nope”

R/ASKMEN DEAD GİVEAWAYS

- She's my wife :)
- "Dated like a 9.5/10. She was a child model but couldn't continue as she got older because she was only like 5'8" or some shit. Blonde, size 0, conventionally attractive in pretty much every way. A lot of it is a ton of fun. People are in awe of your seductive prowess, you can lose arguments and say you're going to go bang your ex-model girlfriend as you walk away from your defeat. Its worth doing once in your life if you can."
- “Hands, beard, shoulders, head, chest”
- "Usually, you are attracted to the person before you find out she is taken. Been in the situation before. You want her to be attracted to you, but at the same time you don't want her to cheat on her dude. "If she cheats with you, she will cheat on you". Ironically, her loyalty attracts you to her even more. (In my case at least). Even worse is when you get mixed signals."
- "Start with a physical change. Get a good haircut. Put some time into grooming yourself. Shave or trim your beard if you’ve got one. Buy some good cologne for when you go out. Start a work out routine and stick to it even when you don’t want to. Doesn’t have to be anything crazy, it could just be sit ups and push ups.“
- "Tickling my wife drives her crazy."

R/ASKMEN LINGUISTIC PATTERNS

- Casual male language:
  - “at it”, “down to”, “got to”, “me and my”, “the ones”, “do something”
- Testosterone/hardo language:
  - “crew”, ”beast”, “built”, “hero”, “bro”
- Impersonal/analytical language:
  - “more that”, “than was”, “perhaps”, “may”, “if so”, “contact with”, “however”
- Language that subtly dis-individualizes women:
  - “the women”, “women that”, “feminists”

R/ASKMEN SIGNATURE FILTH

"Well, most likely if she's [expletive content] she probably likes you. Unless it's for money or drugs, then it's very possible she don't give a [expletive] or even really not like yo ass. Assuming this [expletive] is [expletive content] for free, if she [expletive content] without you having to say shit, that bish in Love, [expletive]. L.O.V.E. If she [expletive content], she might like you and just be an ungrateful rude [expletive], or she think you nasty and therefore your [expletive content]. If That bish don't even let you [expletive content], like right before you [expletive content] she pulls away, dump That [expletive] as soon as [expletive content]. Shit. Not only is she not feelin you, bish was scheming some way to get something besides yo heart out of yo [expletive]. And you don't want an ungrateful rude, prude [expletive], no fun having [expletive content] like Dat anyway. Now I know you might think I'm clownin, but there's a lot of truf to this shit I ain't playin. See for yoself my man. Good luck to ya. Tell her it's [expletive content]. Hopefully it ain't [expletive content] haha even if it is if she like you, or you giving enough dollas, or you tell her she gonna [expletive content]."

WHERE THE MODEL FAİLS

- Homosexuality
  - "My direct report and **my husband** had an emotional/verbal affair. I worked for a company and was friends with lots of people around the office. A guy (24) in another dept wanted something different and we were hiring so he was hired. He was also really good friends with **my husband**. He reported to me and we had a good friendship also. **My husband** and he started running together…"
  - Lesbian women and gay men referencing their significant others leads to strong false predictions. The model only sees "my husband", it can't parse the subtlety.
  - Even subconscious linguistic predictors in the model's calculus may not be useful. It would be interesting to study whether comments by gay men, for example, track better with the male or female subconscious language patterns that we've seen, and whether subject matter is a factor.
- Sarcasm, Imitation
  - "dO yOu EvEn LiFt BrO?", said in AskWomen, graded as highly male-associated. The vectorizer involves lowercasing everything, so the sardonic type style isn't picked up.
- Stereotype Bias
  - "I got a motorcycle."
  - "Off the top of my head, Louis Zamperini's story is pretty fucking insane. Nellie Bly was also fucking hardcore."
  - These are comments by women, but the swearing and motorcycle lead the model in the opposite direction (98% certainty). This is a classic problem in natural language processing We don't want to have models in our applications and commercial infrastructures that reinforce stereotyped perceptions and behaviors.
- Subreddit Pretense
  - "I'm a woman, but…" The 3-gram "I'm a woman" should indicate female easily, but this phrase is actually more likely to be said by a woman replying *in AskMen*. So the model predicts AskMen correctly, but the underlying proxy is flawed
- Wrong for the Right Reasons
  - At its best, the model identifies a woman replying in AskMen (and vice versa) based on linguistic signatures it has learned. There were a handful of examples of this among the incorrect predictions, but the most definitive tend to be off-color.

HEATMAP TEXT IN TABLEAU

- The comment below was highly linked with r/AskMen, but the keywords driving the prediction were not obvious.
- Divergent color-mapping of the text by coefficient in Tableau helps us visualize how the model sees the text.

"You need to be able to distinguish between a girl that likes the attention you're giving her from a girl that likes you. Most girls at a party will smile and talk to you if you approach them with confidence. After you talk to her though, you need to find a way for her to invest in the interaction. If she never invests in it you never really know what she wants. Walking away from the girl without getting any info is a good way to do this. You started the first conversation, now she's on the spot. If she doesn't come up to you and talk to you again, she isn't getting your number. This forces her to be more than just a receiver of attention. If you're leaving the party, or think you might not see her again at that party. Give her your contact info, don't take any of hers. She's then forced to reach out to you if she wants things to go anywhere. Not only will this tell you if she's actually interested in you, it will make her more attracted. She will see that you aren't desperate, and that makes you more attractive. Girls can sense when the only reason you're talking to them is to get their number or something like that. They can see right through you. When you walk away without the number, she realizes you're not just trying to get her number, you actually just wanted to talk to her. That is attractive."

you need to be able to distinguish between a girl that likes the attention youre giving her from a girl that likes you most girls at a party will smile and talk to you if you approach them with confidence after you talk to her though you need to find a way for her to invest in the interaction if she never invests in it you never really know what she wants walking away from the girl without getting any info is a good way to do this you started the first conversation now shes on the spot if she doesnt come up to you and talk to you again she isnt getting your number this forces her to be more than just a receiver of attention if youre leaving the party or think you might not see her again at that party give her your contact info dont take any of hers shes then forced to reach out to you if she wants things to go anywhere not only will this tell you if shes actually interested in you it will make her more attracted she will see that you arent desperate and that makes you more attractive girls can sense when the only reason youre talking to them is to get their number or something like that they can see right through you when you walk away without the number she realizes ...