# The Difference in Language Usage Between Male and Female Online Fashion Communities

Allie Phillips
phillale@umich.edu
University of Michigan

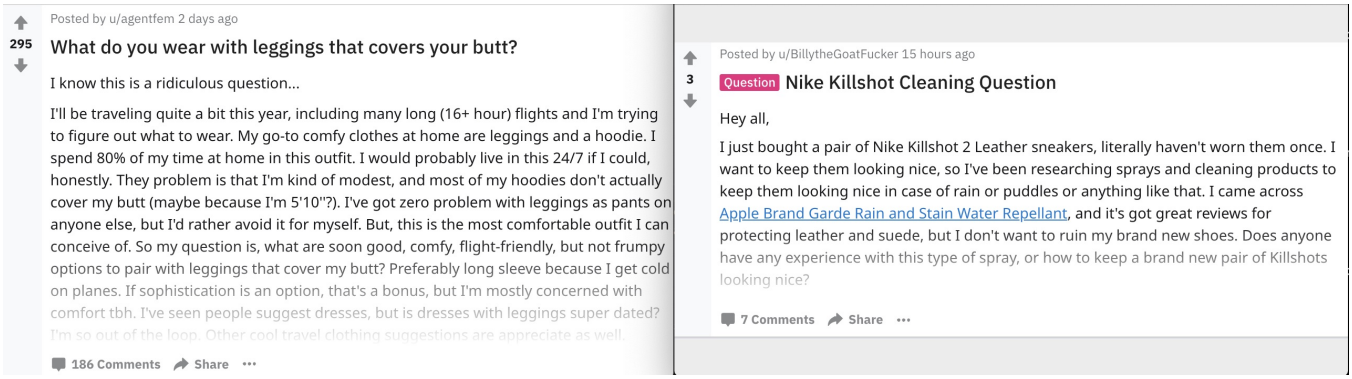


**Figure 1: An example post from femalefashionadvice on the left and an example of malefashionadvice on the right.**

## ABSTRACT

In this research study, I wanted to examine the difference in language use between genders in the fashion communities. To do this, I analyzed comments from the femalefashionadvice subreddit and the malefashionadvice subreddit, as these subreddits are compliments of each other. There are other studies are similar to mine, especially one that was done by Lucy Li and Julia Mendelsohn. In this study, the researchers analyzed gendered communities by looking at the text, the users, and the sentiment. I gathered posts from both subreddits, and narrowed these posts down to a sample size of 10,000 posts. I created a model that would process each post and assign it a value between 0 and 1. If the value was closer to 1, my model would label the post as a post from the femalefashionadvice subreddit. If the value was closer to 0, my model would label the post as a post from the malefashionadvice subreddit. My model performed at 73% accuracy, which indicates that my model was good at deciding which subreddit a post was from. My model also provided results on the most used words in each subreddit that were hardly used in the other subreddit. This information enabled me to make inferences on the difference in language used each subreddit. Finally, I ran an error analysis, where I looked at possible reasons that my model incorrectly labeled posts. From this error analysis it was apparent that gender pronouns had heavy weights, which swayed the way my model labeled the post.

**ACM Reference Format:**
Allie Phillips. 2018. The Difference in Language Usage Between Male and Female Online Fashion Communities. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn


*Conference'17, July 2017, Washington, DC, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

One community that has always been prominent is the fashion community. This tight-knit community has educated its followers on the newest trends, the most popular designers, and the biggest events. Through the internet, this community has been able to grow, as people all over the world have access to fashion magazines, blogs, and TV shows. One interesting question associated with social media fashion communities is whether the rhetoric changes in the way males talk about fashion in relation to the way females discuss fashion. Also interesting is what is most talked about in each subreddit, as the fashion trends seen in online reddit communities may be different than those on a different social media platform. In order to do this, I first gathered data from the malefashionadvice subreddit and the femalefashionadvice subreddit, and created a small random sample of 10,000 posts. I then created a model that was able to read a post and predict which subreddit it was from. Through this process I was able to see the words that heavily related to each subreddit. Using this information, I was able to infer what words were used most often in femalefashionadvice and malefashionadvice. Another accomplishment of my study was being able to train a model and run it on test data.

## 2 RELATED WORK

My research falls under the umbrella of looking at different online gendered communities. On the internet, one has the ability to take on any identity, as other users are unable to see you. This masks one's gender, and provides a platform for anyone to partake in identity play. Though there are many different analyses of online communities, one study in particular is closely related to my research on the different language used in malefashionadvice and femalefashionadvice.

In a 2018 study by Lucy Li and Julia Mendelsohn, researchers analyzed the gendered online communities through three different lenses: text, users, and sentiment. These online communities were represented as different subreddits, which each subreddit containing at least 50,000 subscribers. Once they created a small, random sample of Reddit posts, the researches handpicked users with gendered usernames to analyze their comments [2]. The researchers followed the same methods as I did to figure out the weights of each word in order to determine which subreddit the post is from. From their research, Li and Mendelsohn found that between two communities, word usage does not vary according to a clear cut, binary perspective of gender. They also found that two communities may have high text similarity but not user similarity, or vice versa [2]. When looking at the specific words used, the researchers found that sentiment can be a useful indicator of wordsâĂŹ social meaning and community values.

Overall, the results illustrated that many social media platforms are active settings for exploration of gender identity. In this study, the researchers used the subreddits malefashionadvice and femalefashionadvice. This study is similar to mine, as it looks at the difference in word usage between two subreddits, but it differs in the sense that analyzes more than just two subreddits. It also creates a more specified sample, as it looks at userâĂŹs posts that have gender specific usernames. My work is different than others as it looks at posts from a specific month and the comments used to train my model were randomly selected. Another difference in my project is that I did not take the posterâĂŹs username into account.

In another study, Identifying Misaligned Inter-Group Links and Communities, researchers Srayan Datta, Chanda Phelan, and Eytan Adar looked at the connections between different individuals on social media platforms. The researchers were interested in understanding how inferring what one is saying online determines what implicit links will be created between users [1]. The researchers created an algorithm to identify links, network structures, and communities on social media platforms. The authors then applied this algorithm to Reddit to illustrate inter-group dynamics in social media [1]. This study was similar to mine, as it looked at the language between gendered communities in order to reveal the hidden meaning in the conversations. It also created an algorithm that interpreted what users were saying and provided a method to categorize these conversations. This, however, is where the similarities end. This study is more focused on the implicit and explicit links between people on social media platforms, while my study was more focused on the difference between rhetoric used to discuss fashion in male and female fashion advice subreddits.

## 3 DATASET

The data I am analyzing is everything posted in the malefashionadvice and the femalefashionadvice during December 2018. Getting every single post from this month created a random sample for my model to use, as there were no comments left out. Since the mass amount of comments made my model run slowly, I reduced the amount of comments processed to 10,000. I was still able to maintain a random sample, by first randomizing the comments before narrowing the size down to 10,000.

## 4 METHODS

In order to get data for my model, I first imported the data I received from the subreddits femalefashionadvice and malefashionadvice. To make the data sample easier to run, I had my model create a random sample of 10,000 posts from the two subreddits. From this sample. I divided the 80% of the comments into a training group and 20% of the comments into a testing group. It was important to have 80% of the comments in the training group, as my model needs a large amount of data to train on. I canâĂŹt just train my data on 100% of the data, as I need a test set to make sure that my model actually worked. If I was running the model on my training data, I would get 100% accuracy every time because the model would be memorizing the correct label of the posts. It is important to have a test set, because enables one to see if the model is actually learning how to label the posts. On each set of data, we used the NLTK package to process each word, meaning we took the raw text of the comment and turned it into something easier to understand. Through NLTK, we tokenized the word, made it lowercase, and stemmed it. The reason we used a stemming function was to make sure that different variations of a word were seen as the same value.

Once the NLTK packaged was used on both sets of data, we used a CountVectorizer function to convert the text from each comment into a vector representation that was all the same length. The vector representation used an already determined vocabulary to assign a value to each word depending on whether or not the word was used in the Reddit post. For each word in the vocabulary, if the word was used in the post it would be represented as 1, otherwise it is represented as 0. It also counts how many times the word appears in each sentence.

Once we have done this for every word in the vocabulary, we then have enough information to train our model. We use Logistic Regression, which takes the whole pile of training data, applies SklearnâĂŹs computer magic, looks at the vectors, and spits out an output that gives a set of weights for each post. This squeezing function provides a value between 0 and 1 for each post, representing either femalefashionadvice with a score of one and malefashionadvice with a score of 0. This is how the model learns how to assign a post to one of the subreddits.

This Logistic Regression function takes in the matrix of zeros and ones, and creates a table with the length as the amount of words in a particular set and the width as the size of vocabulary. The size of the vocabulary represents the number of parameters the model has. After running my model in on the training data, I then ran it on my test data which provided me with my modelâĂŹs accuracy score, and a classification report, revealing its recall and precision.

# 5 RESULTS

## 5.1 Model Performance

Looking at the classifier performance, I was able to see how well my model performed. The classifier performance provides results for the precision, accuracy, and recall of my model. The precision and recall of my model are in relation to the positive class, which in my model is femalefashionadvice. This means that if a post got a score that is close to 1, then my model would label this post as being from the femalefashionadvice subreddit. To be labeled as a post from the malefashionadvice subreddit, the post would have to have a score of 0. The precision column represents the total amount of posts that are actually correctly labeled, in relation to the positive class. For malefashionadvice performed with .72 precision, which means that 72% of the time the posts the machine labels as malefashionadvice is actually from malefashionadvice. For femalefashionadvice, the model performed with .74 precision. These results show that the model was more correct when labeling femalefashionadvice posts than malefashionadvice posts. In my model, the accuracy shows the percentage of everything, meaning the true ones and zeros, that the classifier got right. The accuracy of my model was 0.73, which means that my model got 73% of the comments it labeled correct. This number is high, and shows that overall my model performed well. The recall column represents the percentage of true 1âĂŹs the classifier got. The recall for malefashionadvice was 76%, while the recall for femalefashionadvice was 70%. This shows that in relation to the positive class, my model was worse at labeling recognizing the true femalefashionadvice posts. Looking at these percentages, it is apparent that my model was good at deciding whether a post was from femalefashionadvice or malefashionadvice.

# 6 DISCUSSION

## 6.1 Feature Analysis

The output of my model provided the top coefficients in each subreddit, with the top coefficients being the words most used that are not used very much in the other subreddit. This means that these words are so prominent in their subreddit that they are good indicators of which subreddit a post is from. Looking at the most used words, it is apparent that males and females use fashion advice subreddits differently. In femalefashionadvice, the most used words in the subreddit were âĂIJcuteâĂİ, âĂIJskirtâĂİ, âĂIJbraâĂİ, âĂIJbootieâĂİ, and âĂIJloveâĂİ. The use of the word âĂIJcuteâĂİ and âĂIJloveâĂİ would seem to indicate that women use their fashion subreddit for opinions on outfits or individual garments, as these are both describing words. The most used words reveal that âĂIJbraâĂİ and âĂIJskirtâĂİ are the garments that women need the most advice on. The use of these words would seem to indicate that finding a bra or a skirt is more difficult, and once a good bra or skirt is found, women want to share their finds with others. âĂIJBraâĂİ and âĂIJskirtâĂİ also indicate that these items of clothing are the most often searched for by women. The word âĂIJbootieâĂİ would seem to indicate that this is the area of the body that women have the most questions about, as they are mentioning it in an advice subreddit. One can infer that the butt is something that women worry about the most when finding clothes. âĂIJBootieâĂİ also has another meaning, as it could be a type of boot. This may indicate that the most popular type of shoe among women at the moment is the bootie. This can be helpful for retail businesses, as they know what type of shoe they should be advertising in their store.

In the malefashionadvice, the most used words in the subreddit that were not often used in femalefashionadvice were âĂIJmanâĂİ, âĂIJgirlfriendâĂİ, âĂIJdudeâĂİ, âĂIJoliveâĂİ, and âĂIJhoodieâĂİ. âĂIJManâĂİ and âĂIJdudeâĂİ were two words that I was expecting to be used the most in this subreddit, as that is how a lot of males address each other in person. The popularity of these two words demonstrates that males address each other using these words on the internet as well. The use of the word âĂIJhoodieâĂİ would seem to indicate that this is the most popular item of clothing for men, as this is the item of clothing that is talked about the most. One can infer that the hoodie is a staple fashion item for men, as this is the only clothing item in the most used words for the male subreddit. The use of the word âĂIJgirlfriendâĂİ is interesting, as it shows that a manâĂŹs girlfriend is a big topic of conversation in the subreddit. This would seem to indicate that men get a lot of fashion advice from their girlfriends, as they may say in a post, âĂIJMy girlfriend was telling me that this article of clothing is popularâĂİ. Another indication from this word is that men are seeking fashion advice to impress their girlfriends. The popularity of the word âĂIJoliveâĂİ would see to indicate that this is a prominent color in men's fashion. The use of this word is important, as retailers can use this word to predict fashion trends for the coming year. If one retailer knows that olive is popular among males, they can create clothes using the olive color and generate more revenue. It also can be important for retailers targeting women, as they are able to know which colors to avoid. Looking at the most used words in each subreddit, one can gain insight as to what the most popular fashion trends are, or what items of clothing are of most interest to each gender.

## 6.2 Error Analysis

Even though my model performed with 73% accuracy, there is always room for failure. When looking at examples of posts that my model labeled long, it is apparent that there are some patterns in the way my model failed. One major problem my model had was when a post used female pronouns. My model associated female pronouns with highly weighted coefficients, which swayed the way the post was labeled. This led to problems when my model incorrectly labeled posts from the malefashionadvice subreddit as the femalefashionadvice subreddit. This failure exemplifies the high use of female pronouns in the femalefashionadvice subreddit, and indicates that males talk about their female counterparts when discussing fashion in their own subreddit.

Another issue my model ran into was correctly labeling posts that had store names in them. Store names, especially popular department stores, were highly weighted for the femalefashionadvice subreddit. I believe that this reveals it is hard for my model to account for stores that contain both male and female clothing. A final issue that I noticed in my model was difficulty in correctly labeling posts that contained name calling. In one example, a post from the femalefashionadvice subreddit contained the word âĂIJidiotâĂİ. Looking at the coefficient for âĂIJidiotâĂİ indicated that it was heavily associated with the malefashionadvice subreddit. This

suggests that name calling is more prominent in the male subreddit group, and the female users are more kind in their dialogue. The issues that my model ran into demonstrates that this model is susceptible to stereotyping, as it associates females with knowing more store name and males with name calling and harsh language.

## 7 CONCLUSION

In conclusion, I first retrieved data from the femalefashionadvice subreddit and the malefashionadvice subreddit. I narrowed down this data to a random sample of 10,000 comments, which decreased the amount of time it took to run my model. I then trained my model to determine whether a post was from the femalefashionadvice subreddit or the malefashionadvice subreddit. Once my model was trained, I ran it on the test data, which provided me information on the performance of my model. Overall, my model performed well, with an accuracy of 73%. The results of my model provided me model coefficients, from which I was able to derive interesting insights about the language used in the malefashionadvice subreddit and the femalefashionadvice subreddit.

## REFERENCES

[1] Srayan Datta, Chanda Phelan, and Eytan Adar. 2017. Identifying Misaligned Inter-Group Links and Communities. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (Dec. 2017), 37:1–37:23. https://doi.org/10.1145/3134672

[2] Li Lucy and Julia Mendelsohn. 2018. Using Sentiment Induction to Understand Variation in Gendered Online Communities. *arXiv:1811.07061 [cs]* (Nov. 2018). http://arxiv.org/abs/1811.07061 arXiv: 1811.07061.

[2]