



#Twitter

Targeted Advertising



#Meet The Team

Akhila Guttikonda - ag79445



Mihir Ninad Deshpande - md46487



Spoorthi Anupuru - sa56647

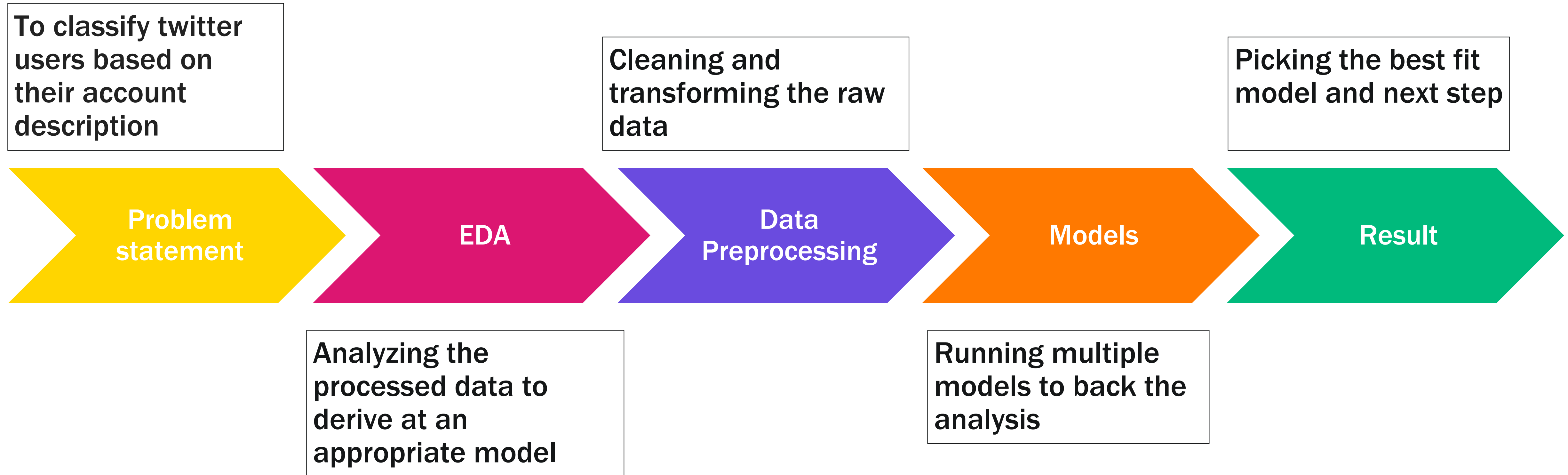


Yvonne Wang - yw22547





#Agenda



#What's Targeted Advertising?



A target market strategy is a business plan that aims to increase sales and brand recognition within a certain demographic of customers. Businesses determine potential target markets based on traits shared by current or potential customers in order to forecast future sales and boost revenues by boosting product sales.

Companies do this through social media posts and ad campaigns.



#What are we trying to do?



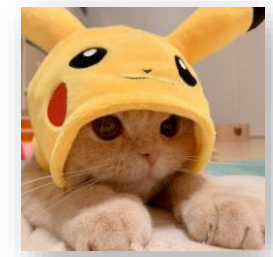
With the increase in the number of social media users and content posted by them, the interest to market products based on demographic factors like age, gender and income by advertising agencies has also increased.

In this project, we have used Twitter database to classify the gender of Twitter users by their profile description to help the advertising agencies for Gender targeted marketing





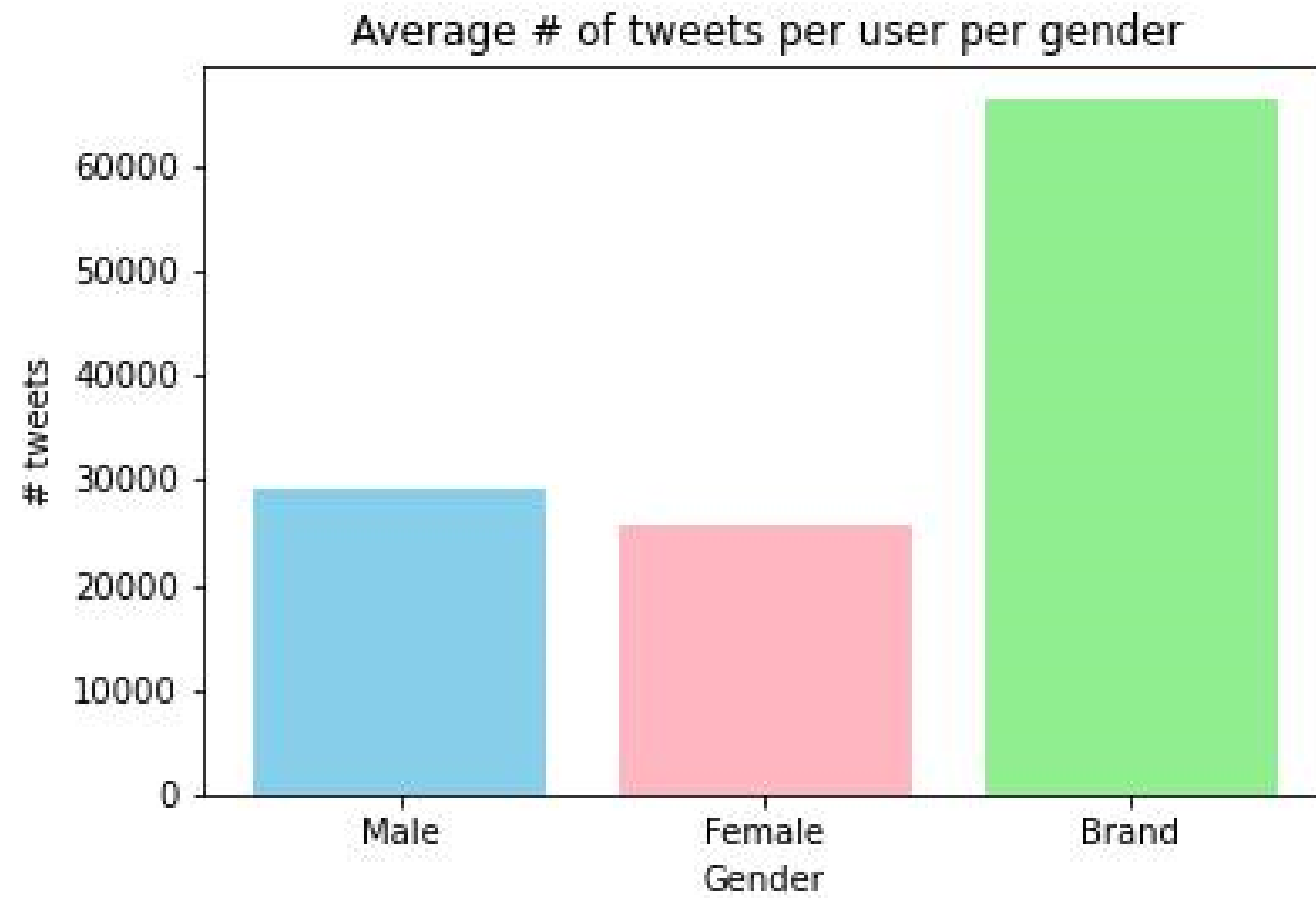
Exploratory Data Analysis



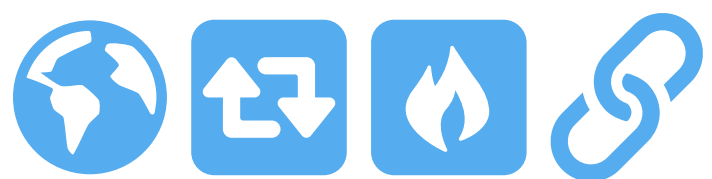
- unitid: a unique id for user
- gender: one of male, female, or brand (for non-human profiles)
- gender:confidence: a float representing confidence in the provided gender
- description: the user's profile description
- fav_number: number of tweets the user has favorited
- name: the user's name
- retweet_count: number of times the user has retweeted (or possibly, been retweeted)
- sidebar_color: color of the profile sidebar, as a hex value
- text: text of a random one of the user's tweets
- tweet_count: number of tweets that the user has posted

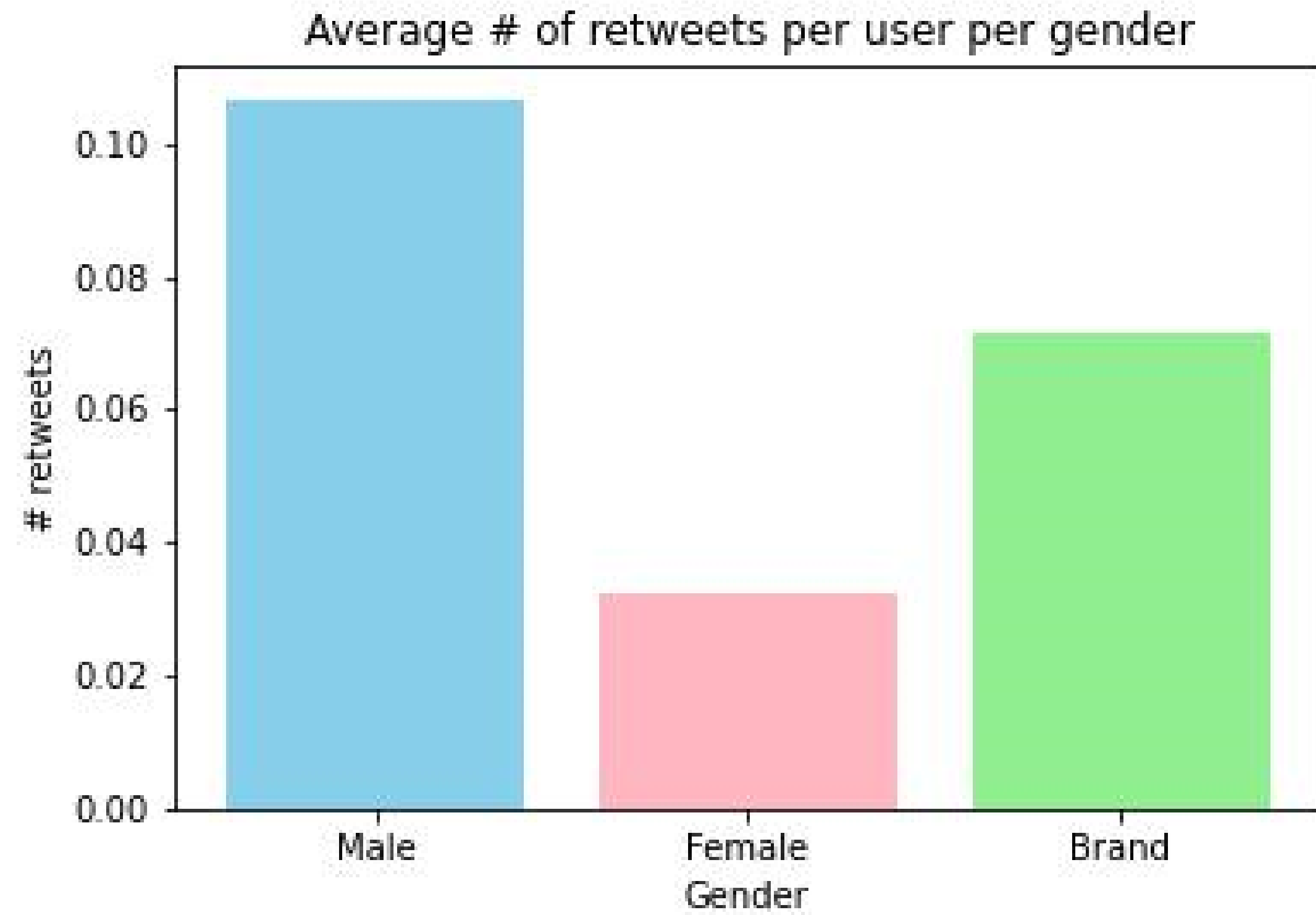
↻ 1M ♥ 2M ↩

Observations = 20,050

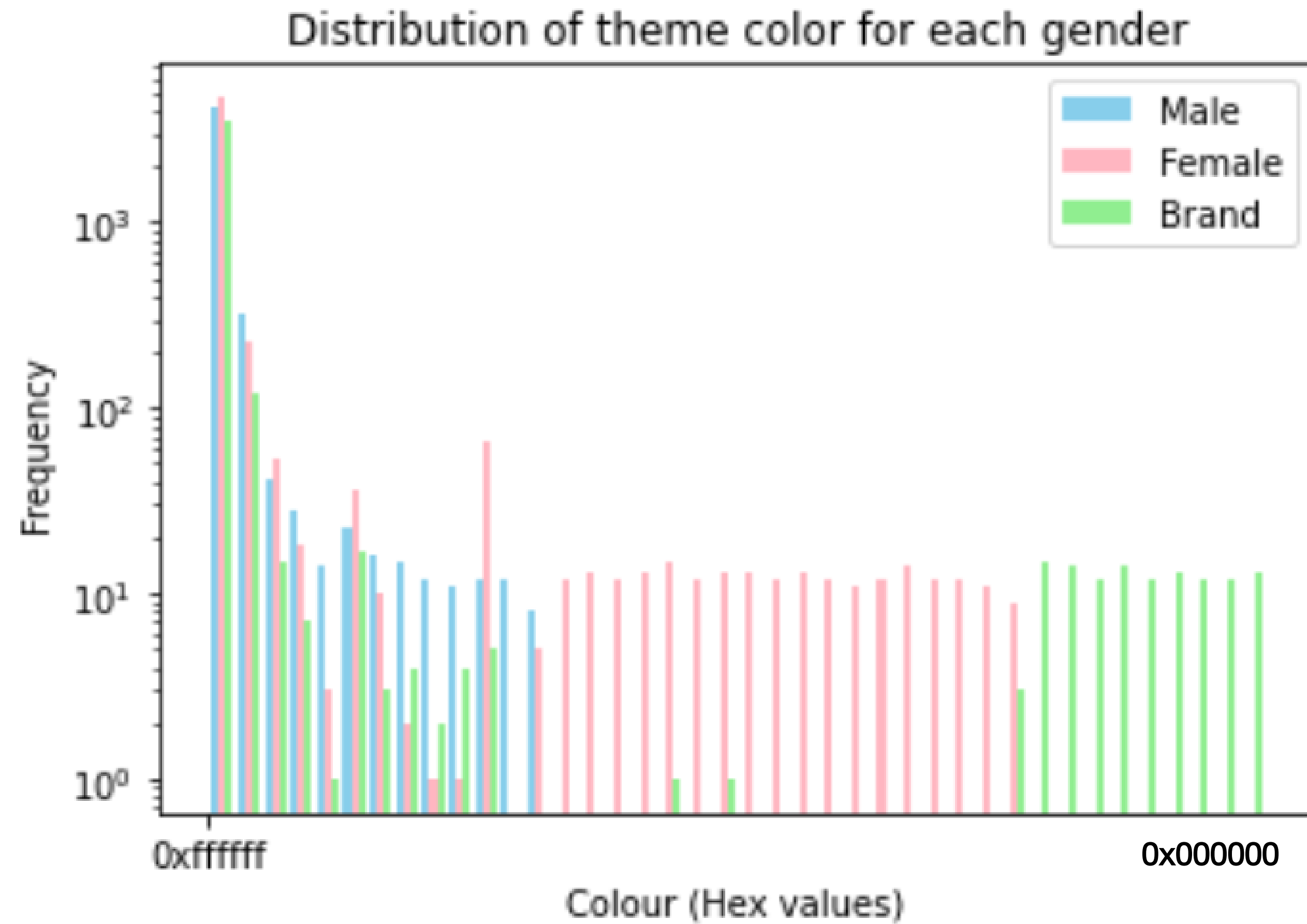


- Brand makes the highest number of tweets
- Male users on average tweet slightly more than Female

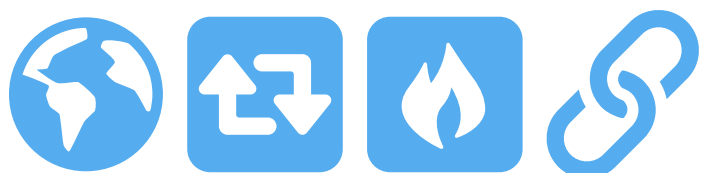




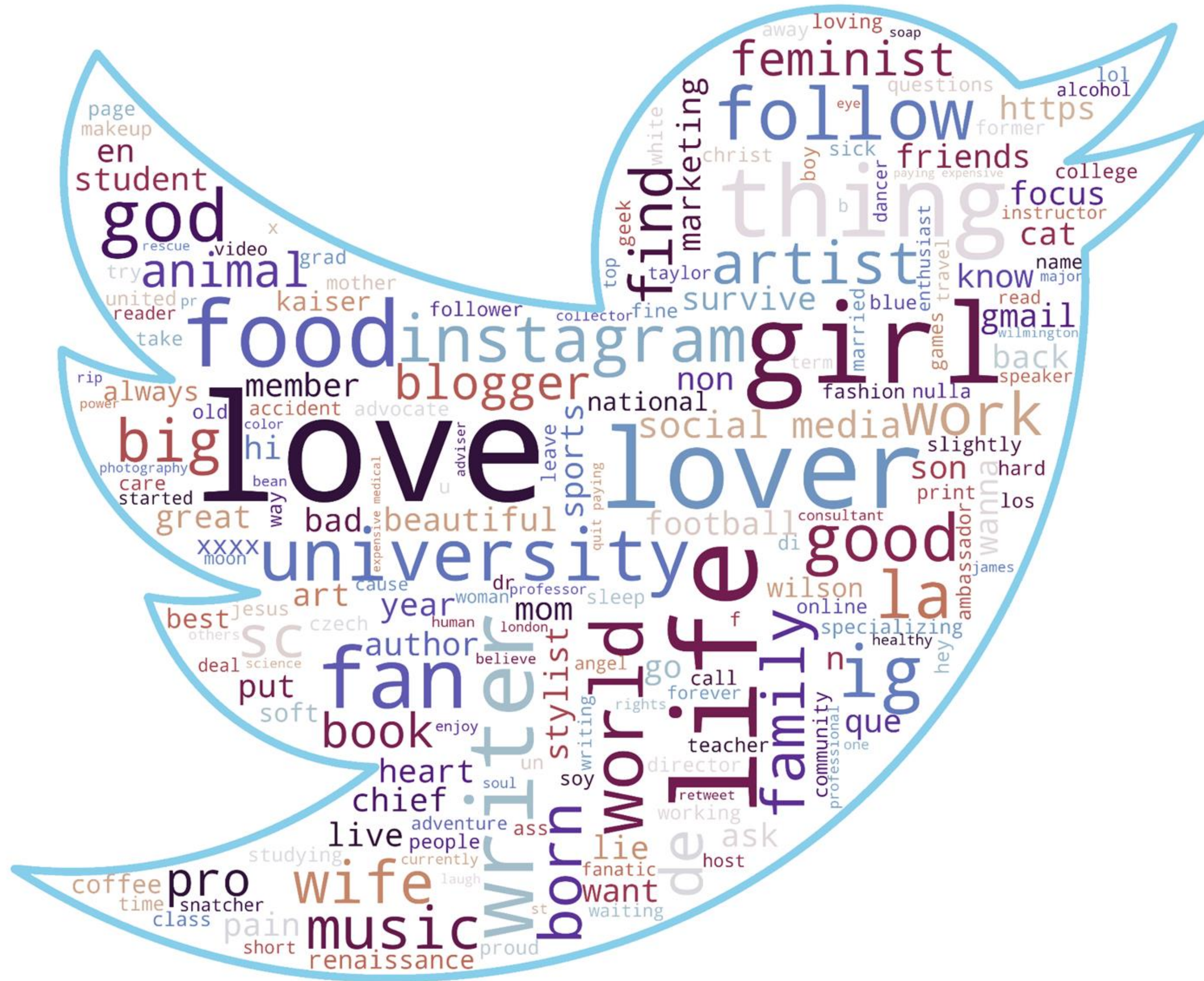
- Male makes the highest number of retweets
- Brand users on average retweet slightly more than female



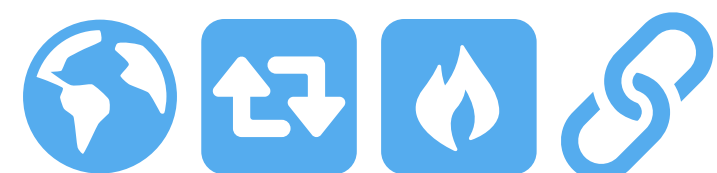
- Most of the users have the default white as their side bar color
- Skewed pattern of sidebar colors wrt gender
- Very little/no inference from the pattern for each gender

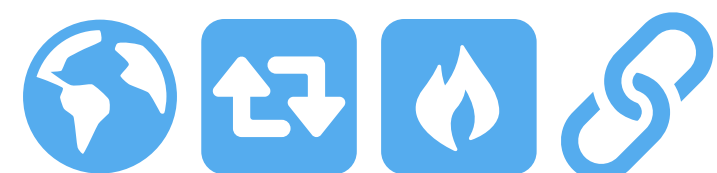
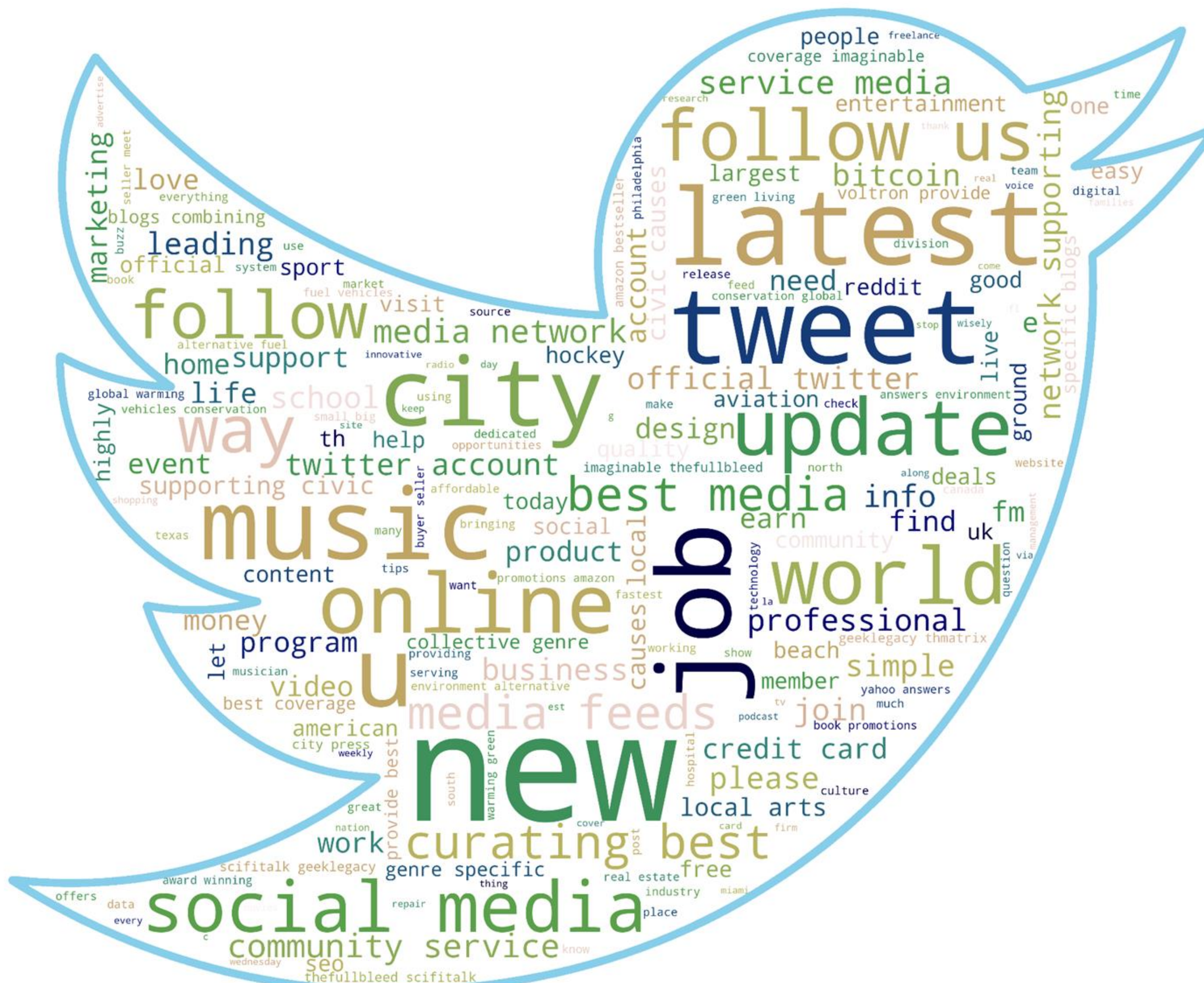


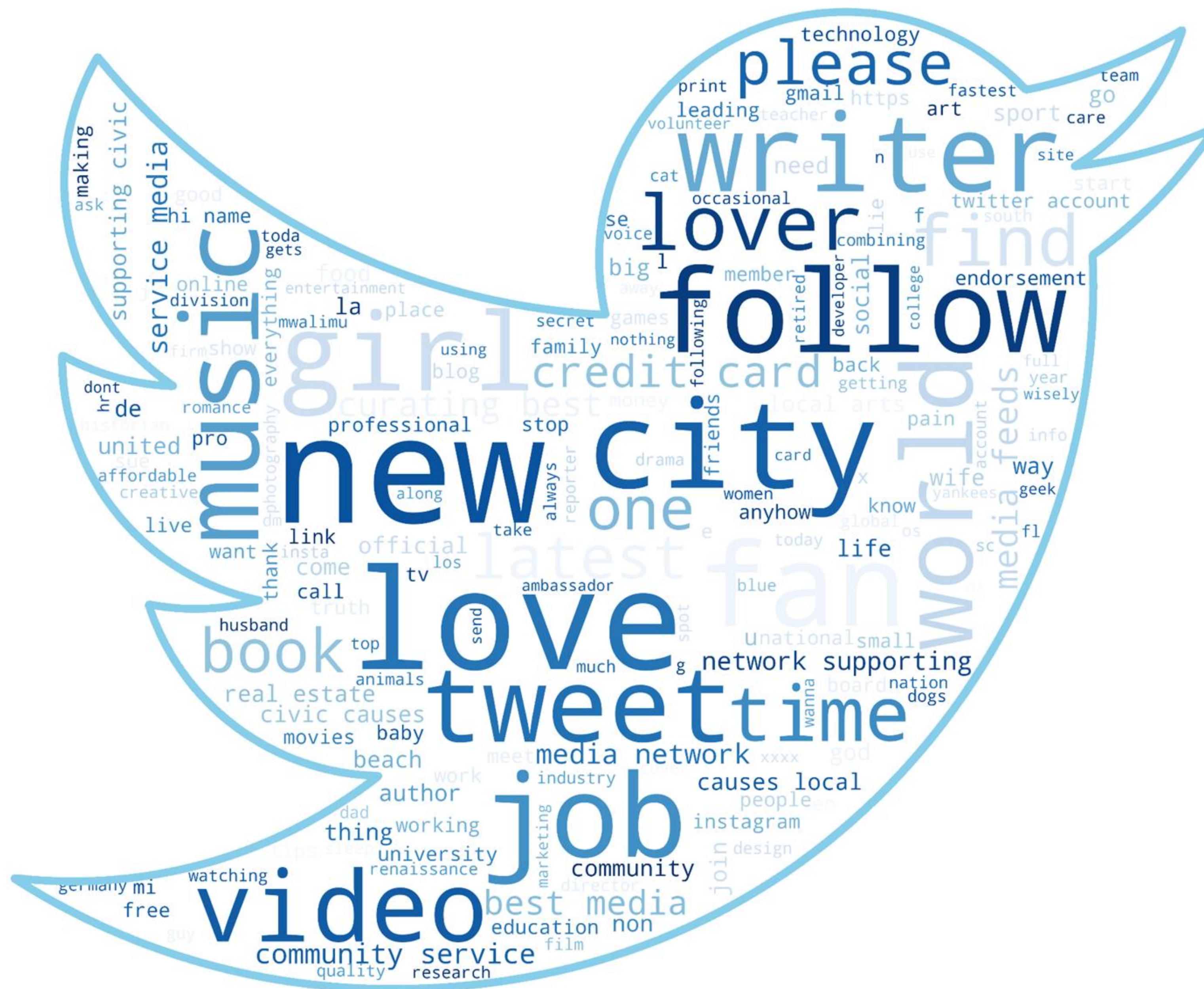




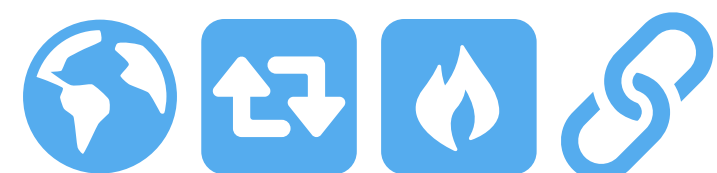
Females







Total





Data Preprocessing



#Cleaning

- Only select entries with 'gender:confidence' = 1
=> Reduced data size to 13,926
- Drop any entries where 'description' is blank
=> Reduced data size to 11,779
- Drop any entries where 'gender' = unknown
=> Reduced data size to 11,773





#Cleaning

- Remove any stop words^[1] from the entire set of descriptions (referred to as corpus)
- Stem^[2] every word from the descriptions to reduce effect of inflections



[1] A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) which is filtered out before or after processing of natural language data because they are insignificant.

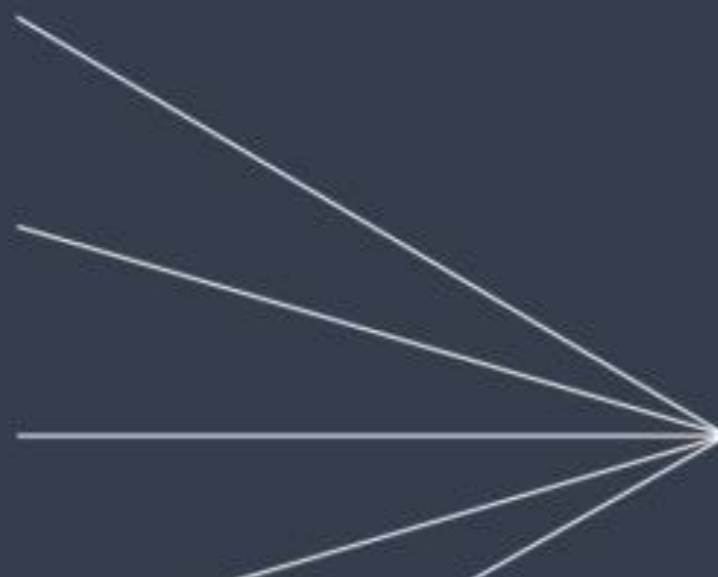
[2] Stemming is a natural language processing technique that lowers inflection in words to their root forms, hence aiding in the preprocessing of text, words, and documents for text normalization.



#Cleaning



connects
connected
connecting
connection
connections



connect



#Cleaning



Before removing stop words and stemming -

```
In [44]: 1 twitter_desc['description'].iloc[1]
```

```
Out[44]: "I'm the author of novels filled with family drama and romance."
```

After removing stop words and stemming -

```
In [49]: 1 corpus[1]
```

```
Out[49]: 'author novel fill famili drama romanc'
```



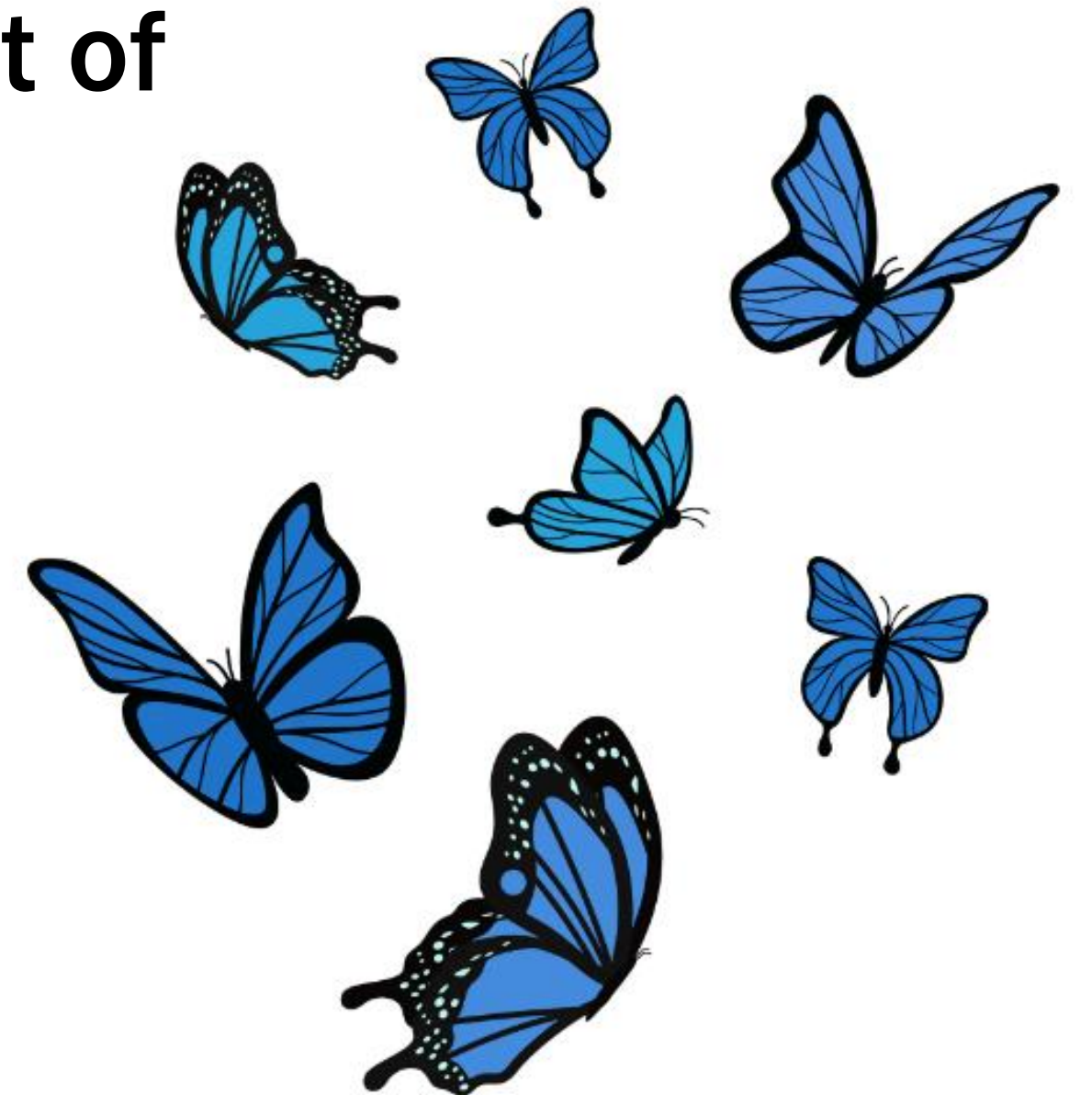


#Transformation



We cannot pass strings (description) as a direct input to any statistical model.

Therefore, we need to transform this string data into a format that can be fed into and analyzed by these models (this is a key part of Natural Language Processing)





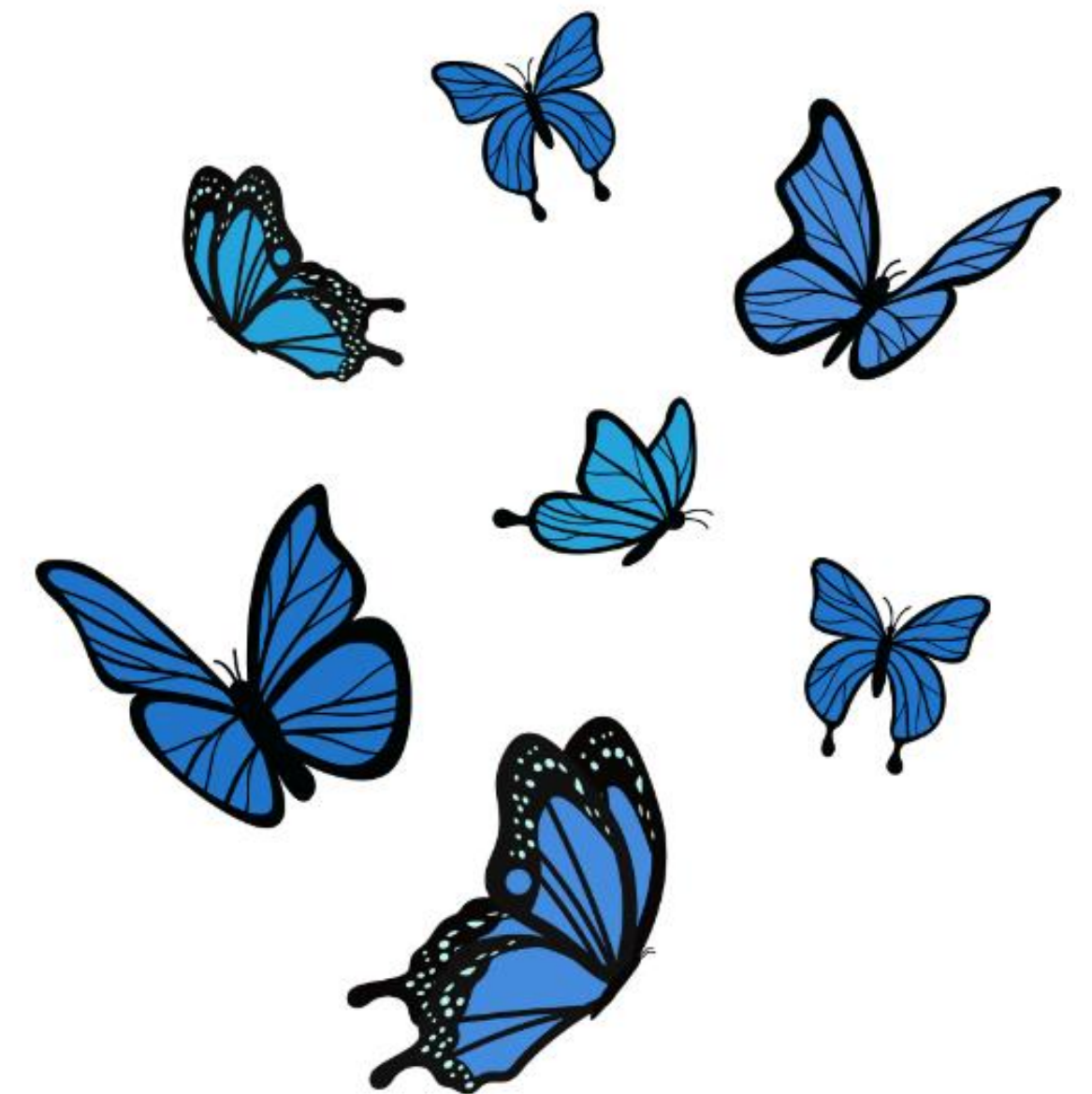
#Transformation



1)Bag of Words

2)N-Grams

3)TF-IDF values





#Bag Of Words

The bag-of-words (BOW) model is a representation that turns arbitrary text into fixed-length vectors by counting how many times each word appears. This process is often referred to as vectorization.

Note that we lose contextual information, e.g. where in the document the word appeared, when we use BOW. It's like a literal bag-of-words: it only tells you what words occur in the document, not where they occurred.



** As per <https://victorzhou.com/blog/bag-of-words/>*



#Bag Of Words

Document	the	cat	sat	in	hat	with
<i>the cat sat</i>	1	1	1	0	0	0
<i>the cat sat in the hat</i>	2	1	1	1	1	0
<i>the cat with the hat</i>	2	1	0	0	1	1



* As per <https://victorzhou.com/blog/bag-of-words/>



#N-grams

An n-gram is a sequence of n words: a 2-gram (which we'll call bigram) is a two-word sequence of words. like “please turn”, “turn your”, or “your homework”, and a 3-gram (a trigram) is a three-word sequence of words like “please turn your”, or “turn your homework”

It is just like the bag of words model, however instead of vectorizing each single word, we vectorize based on sequences of ‘n’ words (bag of words can be considered as a 1-gram model).





#N-grams

Uni-Gram

This	Is	Big	Data	AI	Book
------	----	-----	------	----	------

Bi-Gram

This is	Is Big	Big Data	Data AI	AI Book
---------	--------	----------	---------	---------

Tri-Gram

This is Big	Is Big Data	Big Data AI	Data AI Book
-------------	-------------	-------------	--------------



** As per Speech and Language Processing. Daniel Jurafsky & James H. Martin, Stanford*



#TF-IDF

TF-IDF stands for “Term Frequency — Inverse Document Frequency”. This is a technique to quantify words in a set of documents. We generally compute a score for each word to signify its importance in the document and corpus. This method is a widely used technique in Information Retrieval and Text Mining

TF-IDF vectorization involves calculating the TF-IDF score for every word in your corpus relative to that document and then putting that information into a vector



** As per <https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/>*



#TF-IDF

Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	$\log(2/2) = 0$	0	0
Car	1/7	0	$\log(2/1) = 0.3$	0.043	0
Truck	0	1/7	$\log(2/1) = 0.3$	0	0.043
Is	1/7	1/7	$\log(2/2) = 0$	0	0
Driven	1/7	1/7	$\log(2/2) = 0$	0	0
On	1/7	1/7	$\log(2/2) = 0$	0	0
The	1/7	1/7	$\log(2/2) = 0$	0	0
Road	1/7	0	$\log(2/1) = 0.3$	0.043	0
Highway	0	1/7	$\log(2/1) = 0.3$	0	0.043

Models





#Baseline Accuracy

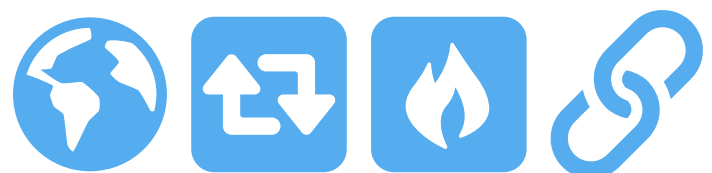
Males = 4150

Females = 4616 <- Maximum

Brands = 3007

Baseline Accuracy = $4616/11773 = 39.21\%$

This is the accuracy we need to beat!





#Multinomial Naive Bayes

Multinomial Naive Bayes classification algorithm tends to be a baseline solution for sentiment analysis task. The basic idea of Naive Bayes technique is to find the probabilities of classes assigned to texts by using the joint probabilities of words and classes

$$P(C_k | x_1, \dots, x_n) = \frac{P(C_k)P(x_1, \dots, x_n | C_k)}{P(x_1, \dots, x_n)}$$

* As per <https://towardsdatascience.com/sentiment-analysis-of-tweets-using-multinomial-naive-bayes-1009ed24276b>





#Multinomial Naive Bayes

Bag of Words

-4.0986 http
-4.2028 news
-4.9832 follow
-5.0816 us
-5.1108 tweet
-5.3579 twitter
-5.3731 offici
-5.4124 world
-5.4285 busi
-5.4616 servic
-5.4957 updat
-5.5133 latest
-5.5493 help
-5.5678 best
-5.5772 free
-5.5963 account
-5.6660 music
-5.6763 new
-5.6868 provid
-5.7298 game

N-Grams

(N=2)
-5.2169 offici twitter
-5.3992 follow us
-5.4747 price updat
-5.4747 continu price
-5.6224 visit http
-5.6930 latest news
-5.9101 gmail com
-5.9408 twitter account
-6.1107 tweet us
-6.1107 need help
-6.1107 improv health
-6.1485 spiritu empow
-6.1485 secur improv
-6.1485 research prove
-6.1485 prove visit
-6.1485 http grfoxfjwpv
-6.1485 health research
-6.1485 financi secur
-6.1485 empow financi
-6.1877 social media

TF-IDF

-4.8492 news
-5.6596 us
-5.6614 follow
-5.6693 updat
-5.7876 tweet
-5.7887 offici
-5.8431 twitter
-5.8588 latest
-6.0616 world
-6.0656 servic
-6.0710 price
-6.0756 account
-6.1177 continu
-6.1285 best
-6.1293 busi
-6.1617 help
-6.2061 new
-6.2628 free
-6.2792 provid
-6.2883 game

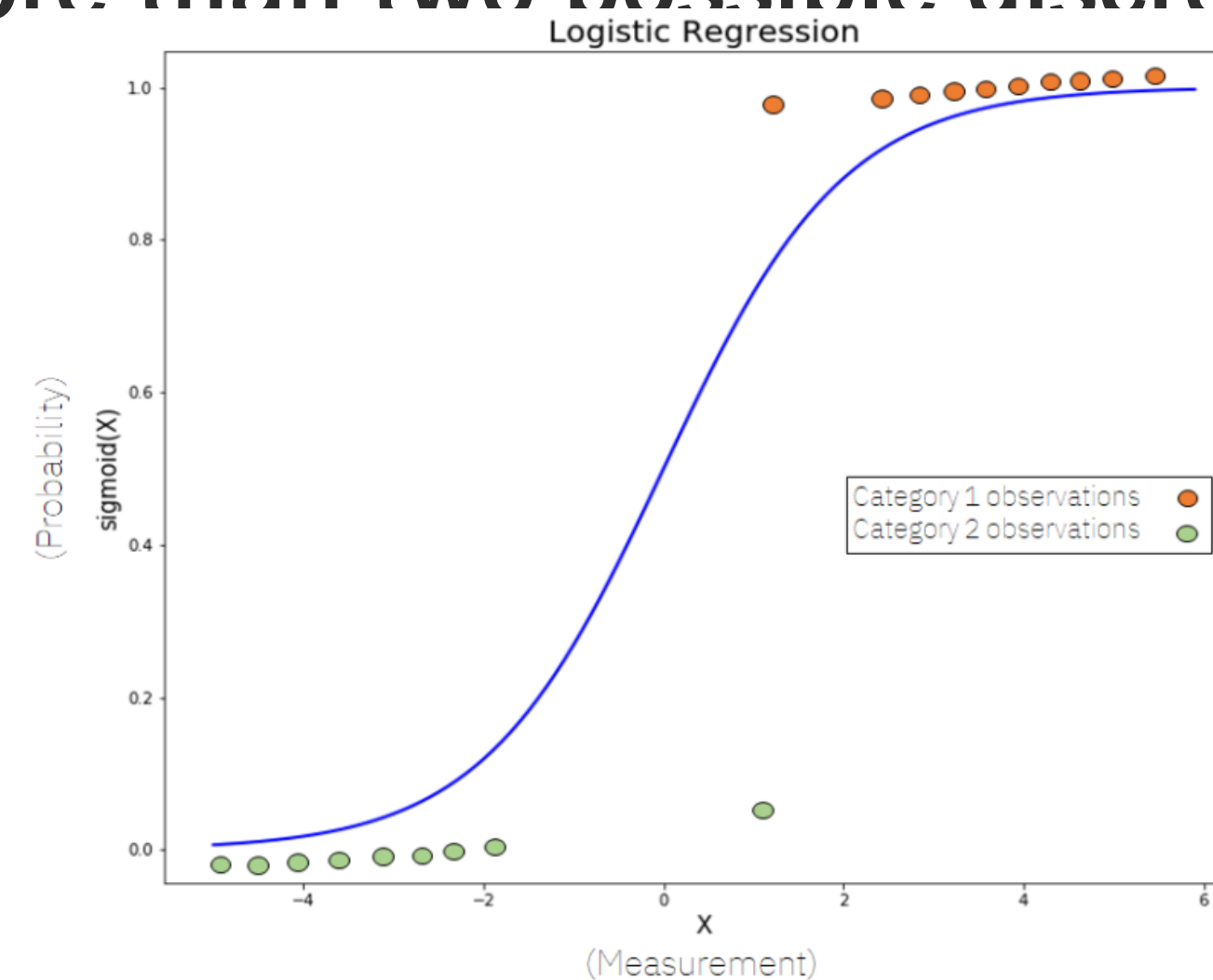
Standard 80-20 train/validation split to check for Out Of Sample (OOS) accuracy





#Logistic Regression

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression model is a binary outcome; something that can take two values such as true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes



* As per <https://www.sciencedirect.com/topics/computer-science/logistic-regression>





#Logistic Regression

Bag of Words

2.2699	weather
1.6088	news
1.5011	us
1.4873	restaur
1.4401	jewelri
1.4211	organ
1.4204	resourc
1.4186	station
1.3816	offici
1.3790	continu
1.3767	inform
1.3614	reddit
1.3568	discov
1.3540	worldwid
1.3344	locat
1.3239	program
1.3113	bot
1.3046	auction
1.2976	buzz
1.2490	updat

N-Grams

2.3242	follow us
1.8567	real time
1.8132	news reddit
1.8092	high qualiti
1.7395	latest news
1.7049	visit http
1.6649	non profit
1.6173	price updat
1.6173	continu price
1.5488	join us
1.5365	break news
1.5076	offici twitter
1.4382	one place
1.4131	custom servic
1.4017	jewelri design
1.3895	news updat
1.3884	check http
1.3817	news inform
1.3751	offici account
1.3551	radio station

TF-IDF

4.7117	news
3.2619	us
2.6869	offici
2.5676	servic
2.3902	inform
2.3762	updat
2.1770	provid
2.0413	visit
2.0398	weather
1.9917	locat
1.9156	help
1.8949	latest
1.8881	station
1.8543	websit
1.8436	price
1.7326	event
1.7079	program
1.6955	shop
1.6882	organ
1.6844	bring

Standard 80-20 train/validation split to check for Out Of Sample (OOS) accuracy



#Extreme Gradient Boosting (XGBoost)



XGBoost is an ensemble learning algorithm meaning that it combines the results of many models, called base learners to make a prediction. Just like in Random Forests, XGBoost uses Decision Trees as base learners.



** As per <https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390>*

#Extreme Gradient Boosting (XGBoost)



The trees used by XGBoost are a bit different than traditional decision trees. They are called CART trees (Classification and Regression trees) and instead of containing a single decision in each “leaf” node, they contain real-value scores of whether an instance belongs to a group

After the tree reaches max depth, the decision can be made by converting the scores into categories using a certain threshold



* As per <https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390>

#Extreme Gradient Boosting (XGBoost)



1. `learning_rate`: also called *eta*, it specifies how quickly the model fits the residual errors by using additional base learners.

- typical values: 0.01–0.2

2. `max_depth` - how deep the tree's decision nodes can go. Must be a positive integer

- typical values: 1–10

3. `subsample` - fraction of the training set that can be used to train each tree. If this value is low, it may lead to underfitting or if it is too high, it may lead to overfitting

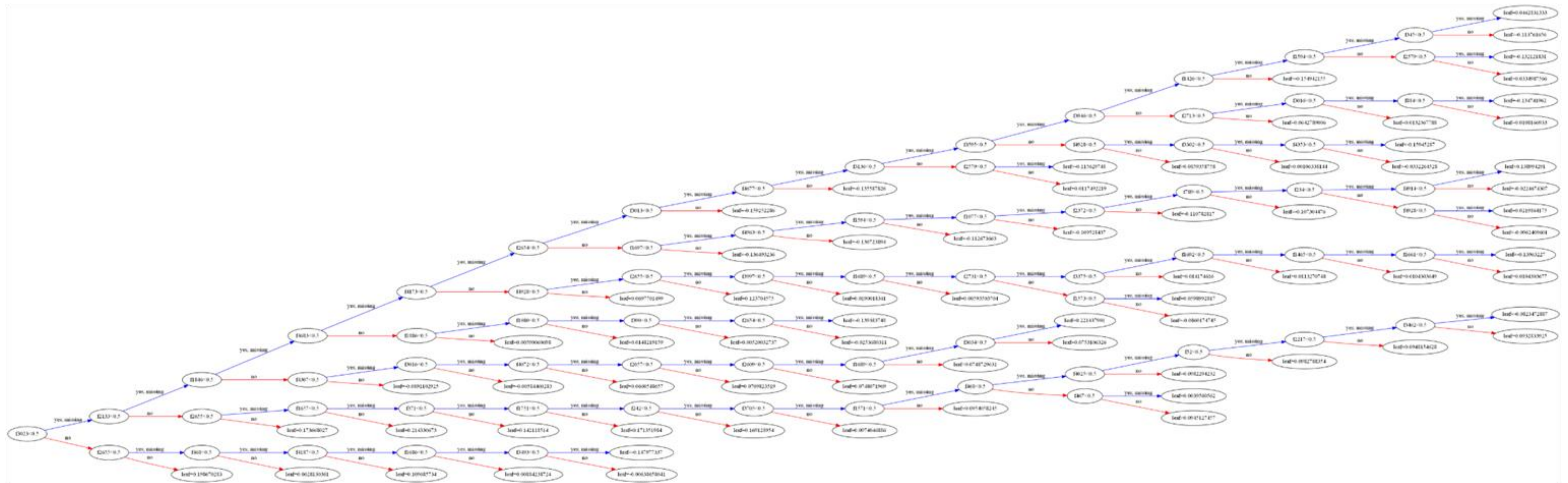
- typical values: 0.5–0.9

4. `n_estimators`: Number of boosting rounds.

* As per <https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390>



#Extreme Gradient Boosting (XGBoost)



* As per <https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390>





Results



#Accuracies

Model	Multinomial Naïve Bayes	Logistic Regression	XG Boost
Bag of Words	65.77%	64.16%	64.79%
Bigrams	54.05%	54.31%	49.05%
TF IDF	65.43%	64.62%	64.16%



#Next Steps

Although we achieved a 65.77 percent accuracy with BOW, there is still scope to improve the accuracy using other techniques, such as using different vectorization methods like GloVe, Word2Vec, etc or a completely different approach such as using Convolutional Neural Networks to predict the gender based on profile pictures of users.





Conclusion



#Conclusion

- Because social media dominates the marketing industry in the modern era, this analysis would be more helpful
- Therefore, by making use of all of these strategies, we can help businesses advertise by choosing the ideal target market for the right goods to enhance their sales







Thank you!