Twitter & Gender

Team 19: The Tweeps

D.J. Anderson, Nishant Jha, Edlene Miguel, Kirsten White

Problem Definition and Background

Our project aims to use Twitter metadata to accurately determine and classify a user's gender. Today, many interactions occur online between users who have limited information on each other or who may never meet in person. For example, this applies to online dating and selling or purchasing items online. In any of these situations, being able to verify and trust the user on the other side of the screen is extremely useful. User gender classification is also helpful from a business analytics perspective, as selling targeted ads to users is a lucrative problem.

There can be some unpleasant ramifications of making predictions like this, though. Twitter added a "gender" field to its user profiles recently, and in May of 2017 many users noticed that Twitter had guessed their gender if one had not been supplied. Some were upset that the classification algorithms seemed to only yield "male" or "female" results and nothing more nuanced. Others made complaints that Twitter was implementing this algorithmic speculation instead of working on solving the many harassment-related problems faced by users of the platform.

Related Work

Some related work that used machine learning to classify a Twitter user's gender includes "Gender Recognition Algorithm for Social Media: Twitter case" by Fi-Ware Consoft and "Detecting the Gender of a Tweet Sender" by the University of Regina's Department of Computer Science. The former article used decision trees and what they called a "waterfall evaluation" to classify such attributes as screen name, description, and profile color. The latter used a support vector machine along with ranker-filter algorithms they obtained from the WEKA Toolkit. The latter also used feature selection algorithms, including chi square, information gain ratio, relief, and symmetrical uncertainty. They were both able to achieve a gender-classifying accuracy of about 87%. Our project is different in that we parse an individual's tweet, in addition to using some of the other aforementioned features like profile color, in order to classify the author's gender.

Description of Data

The data for this project is taken from Kaggle's Twitter User Gender Classification Dataset. Data that is especially valuable for this project are the user data (name, user description, profile image, etc.) and tweet data (text, location, timezone, etc.). The dataset also includes gender classification data, which lists the predicted gender of the user (male/female/brand), the confidence value (how confident they are that the classification is correct), and the unit state (if the classification has been confirmed by humans).

Preprocessing

Several preprocessing steps had to be completed before using the data in our algorithms. Common words like “a,” “the,” “or,” and “be” were removed from the contents of each tweet because those are words that both genders are very likely to use similarly. This means that those words make very little, if any, contribution to determining the author’s gender. Other non-alphabetic characters, including numbers, ampersands, punctuation, non-English characters (letters with accents), and emojis were also removed from the contents of the tweet. These characters are either used infrequently or used among both genders with approximately equal proportion and depend on other factors like the author’s race and ethnicity rather than strictly gender. In addition, we also identified the twenty most common words and bigrams used in tweets, descriptions, and the combination thereof by each gender. This preprocessing step of 20,050 tweets took between two and three seconds. Figures 1-18 in the appendix show the most common male, female, and brand words and bigrams among tweets and descriptions.

One additional preprocessing step we performed was to compile a list of the words used exclusively by any of the genders. We decided that using this data as one of our primary features would quite likely lead to overfitting, and therefore did not use it in our decision making process. This preprocessing step was very inefficient, taking just over 20 minutes to complete. For the curious reader, these lists of gender-exclusive words are included in the attached file, *exclusive.txt*.

Feature Selection

After preprocessing, the most significant portion of feature extraction and selection was related to the individual words used in tweets and descriptions. We decided to compute the “maleness,” “femaleness,” and “brandness” of tweets and descriptions. This was done by first computing the percentage of tweets and descriptions from each gender that included each of the top 2000 words used by each gender. Next, those values were normalized by subtracting the percentage of the other two genders’ tweets and descriptions that also contained that word. This yielded the “maleness,” “femaleness,” and “brandness” of each word. Finally, the average “maleness,” “femaleness,” and “brandness” of each tweet and description were found.

To further reduce the dimensionality of our data, we intuitively removed fields that seemed like they would have little to no correlation with the gender of the user, such as user location, tweet coordinates, date/time of account creation, and date/time of tweet creation. The features that we decided were among the most important were the user’s description, link color, sidebar color, number of tweets, and number of favorites, and the tweet’s content and number of retweets.

Experimental Design

Process

The overall process we followed to perform this study went as follows. We began by preprocessing the data. Next, we performed feature extraction and selection. Once that was complete, we executed two different algorithms on the resulting data. For the decision tree, we performed 4-fold cross validation, and for kNN, we performed 10-fold cross-validation. After the algorithm execution, we compared various metrics of the two algorithms to determine which was more successful in helping us identify the gender of a Twitter user.

Algorithms

Decision Trees

We chose to use the ID3 algorithm and information gain to determine which features are most important and helpful in determining a user's gender. Since the data is split at every node, we might reach a node where there is not enough data to carry out a classification, which would indicate an overfitting of the data.

K-Nearest Neighbor

We chose to use the kNN algorithm to help us solve this problem because it seemed as though similar tweets and/or users would likely end up being clustered together, and using the nearest neighbors of a tweet to "vote" on its gender could therefore be an accurate method of classification.

Model Selection

We used k-fold cross validation in our implementations of the decision tree and kNN to minimize the likelihood of overfitting and to determine how well the model generalizes. For the decision tree we chose k = 4, and for kNN we chose k = 10.

Metrics

We used confusion matrices to determine which model was more accurate at predicting gender by representing the results of the predictions alongside the results of the actual class. For the decision tree approach, which uses a 4-fold cross validation implementation, each of the 4 resulting trees has its own confusion matrix. From the confusion matrices, we compute the accuracy of each tree. We select the most accurate tree as the best of the four folds. For the kNN approach, which uses a 10-fold cross validation implementation, we evaluate the accuracy of our results by comparing MSE values of the 10 resulting folds to determine which fold has the lowest misclassification error.

Results and Conclusions

Because we were deciding between three different classifications, a random decider would be expected to have an accuracy rate of roughly 33% (instead of 50% in the case of a binary classification).

Our kNN implementation demonstrated that an optimal number of neighbors is 33. The accuracy of that classifier was 48%. This indicates that it performed better than random, but was not incredibly reliable.

Our decision tree implementation returned accuracies of 51.7%, 51.3%, 50.6%, and 51.4% for each of the folds, respectively. The maximum of these was 51.7% and the average was 51.3%. Similar to the kNN implementation, this demonstrates an improved performance over a random guesser but is still far from ideal.

Overall, we experienced mild success with our implementations of machine learning algorithms, but could have seen marked improvement upon that success had more features been included. Some examples of these features are discussed in the “Future Work” section later in this paper.

Software and Hardware Used

We used various software packages in this study, including Anaconda Notebook’s Jupyter and Spyder platforms, the pandas data analysis library, scikit-learn’s tools for machine learning, and matplotlib data plotting library for data visualization. We each used our own personal laptop computers to carry out the computation.

Future Work

For future work, we think that the inclusion of the “@” symbol, which is included in tweets where users are either starting a conversation or replying to another user, could be indicative of gender. Perhaps males are more likely than females to engage directly with other users. By the same token, analytic inclusion of the hashtag (“#”) could also reveal some trend related to gender. It may be that females include hashtags in their tweets more often than other users do. It would also be wise to consider emojis. Personal experience and anecdotal observations show that there are some emojis that are used by one gender much more than the other (for instance, women use the heart emoji frequently and men are less inclined to use emojis overall). If these were explicitly included in the decision-making process, it would likely yield more accurate models.

One direction of future work that we intended to address in our study was the analysis of bigrams. However, this proved too taxing on our fairly limited computational resources and was therefore omitted from our decision algorithms.

References

Crowdflower. Twitter User Gender Classification | Kaggle, 21 Nov. 2016, www.kaggle.com/crowdflower/twitter-user-gender-classification.

Notopoulos, Katie. “Twitter Has Been Guessing Your Gender And People Are Pissed.”BuzzFeed, www.buzzfeed.com/katienotopoulos/twitter-has-been-guessing-your-gender-and-people-are-pissed?utm_term=.sfe90R3MnL#.ovGDpX9j7o.

“Classification And Regression Trees for Machine Learning.” Machine Learning Mastery, 20 Sept. 2017, machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/.

https://forge.fiware.org/docman/view.php/47/5275/GenderRecognitionAlgorithmGRA.pdf

<u>Appendix</u>

Figure 1: Brand Bigrams in Descriptions

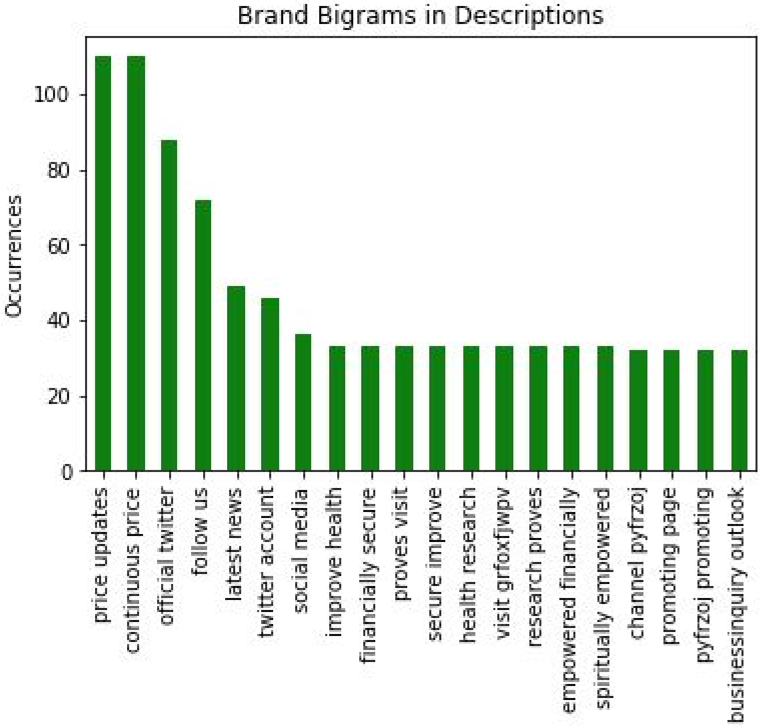


Figure 2: Brand Words in Descriptions

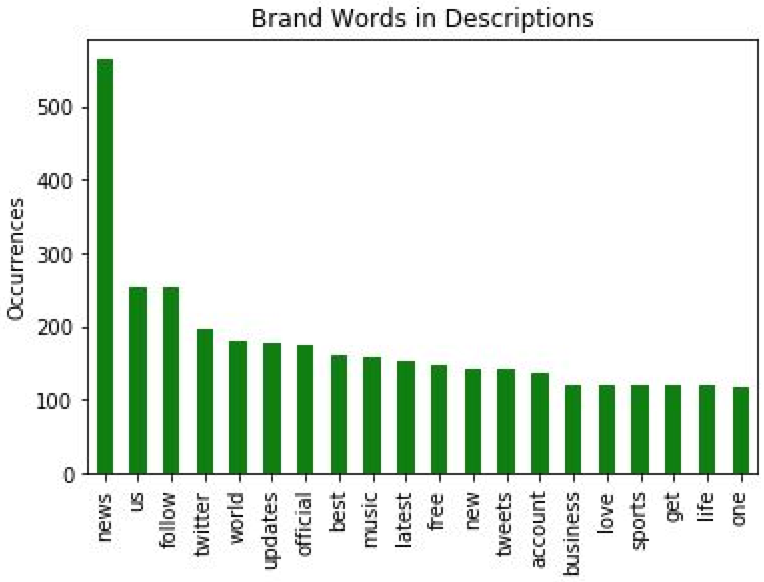


Figure 3: Brand Bigrams Overall

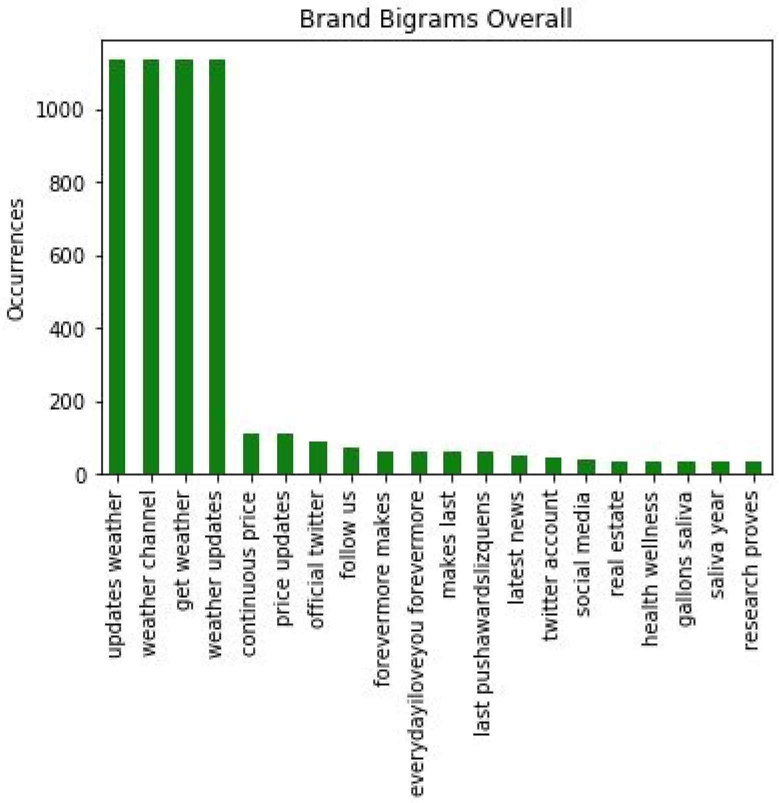


Figure 4: Brand Words Overall

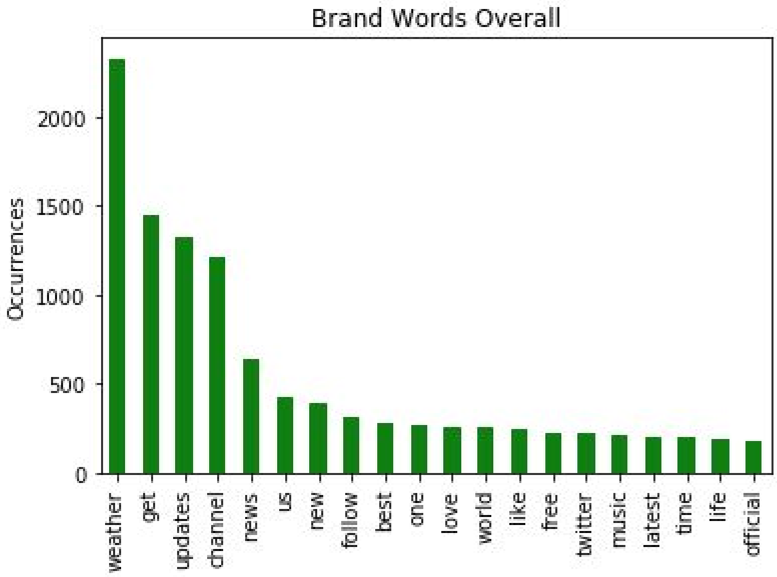


Figure 5: Brand Bigrams in Tweets

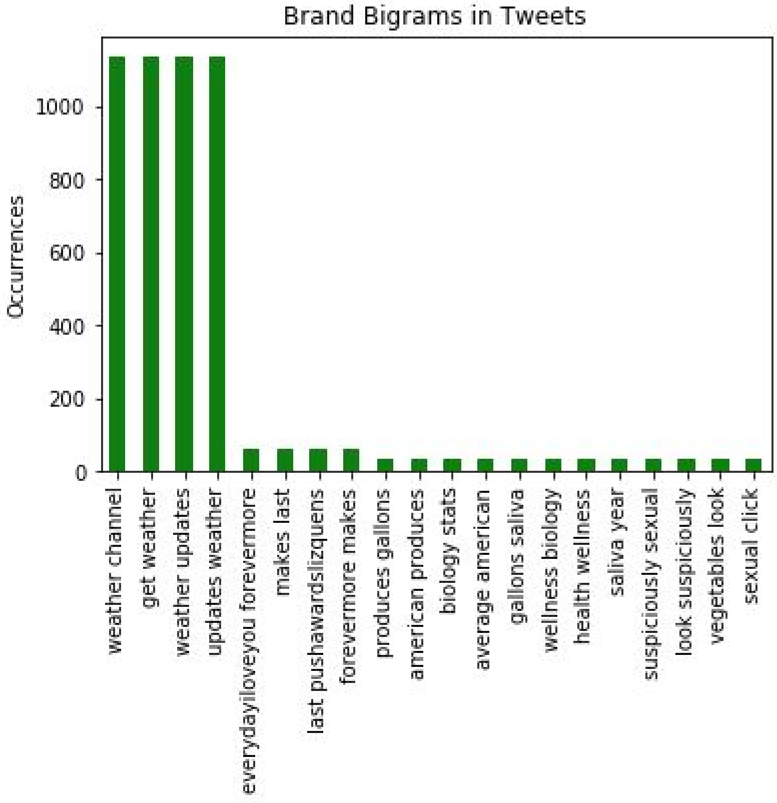


Figure 6: Brand Words in Tweets

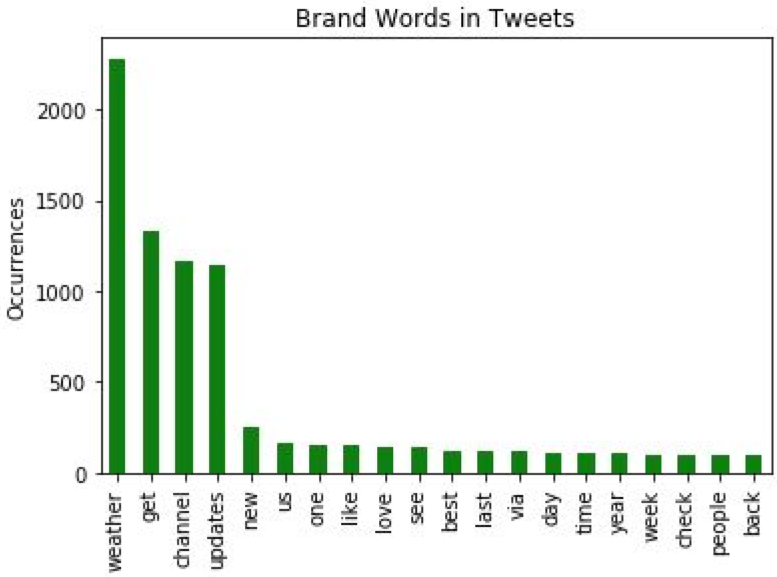


Figure 7: Female Bigrams in Descriptions

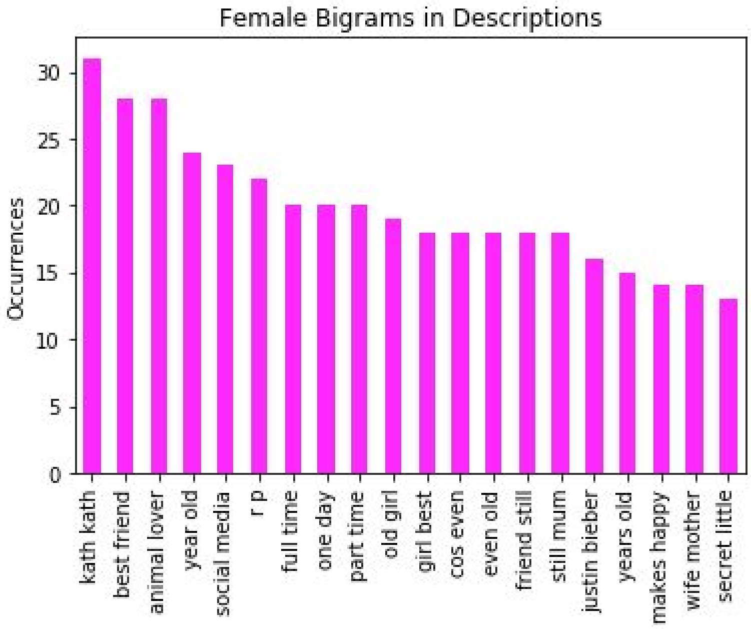


Figure 8: Female Words in Descriptions

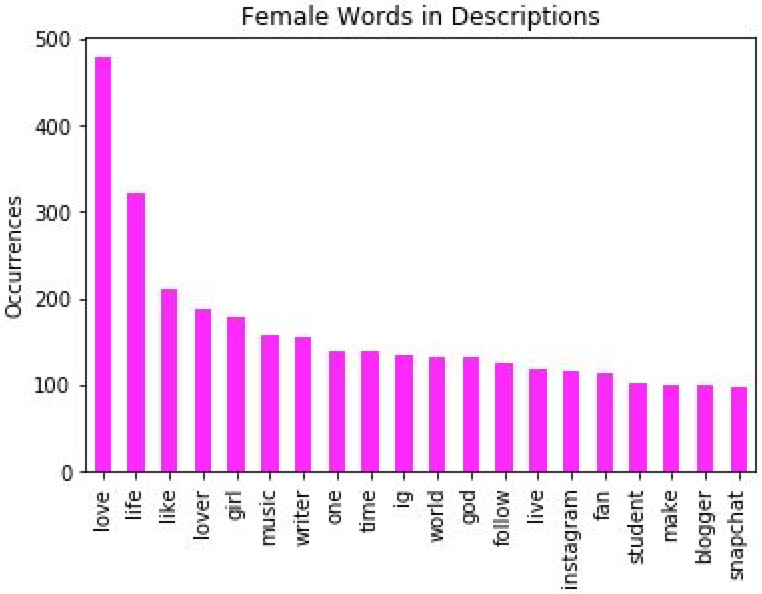


Figure 9: Female Bigrams Overall

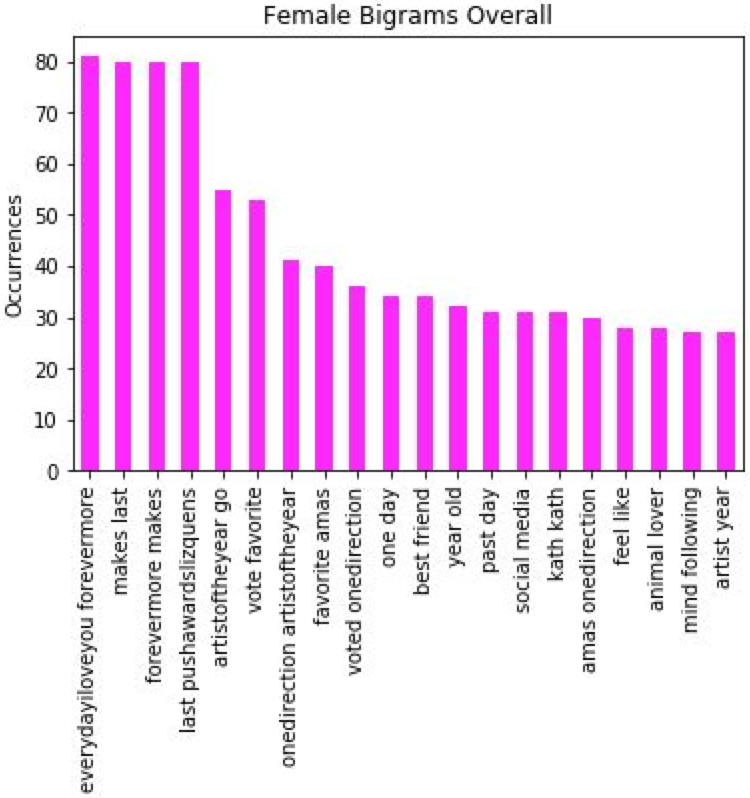


Figure 10: Female Words Overall

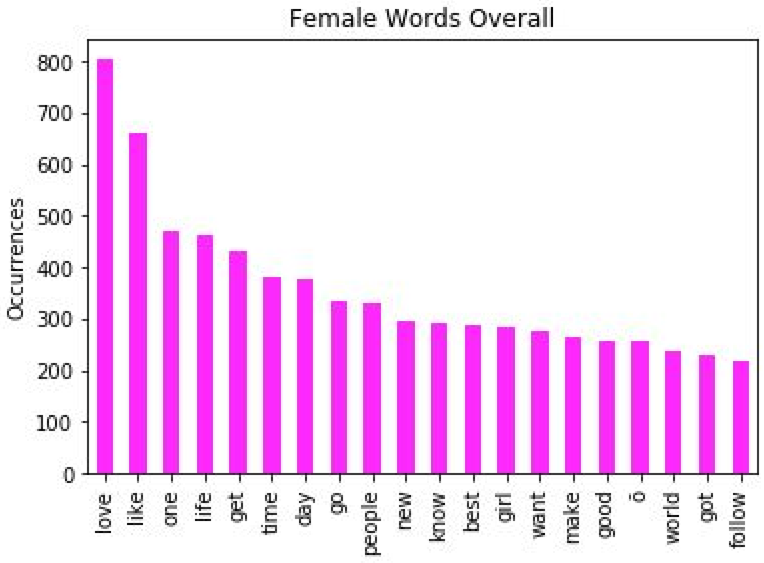


Figure 11: Female Bigrams in Tweets

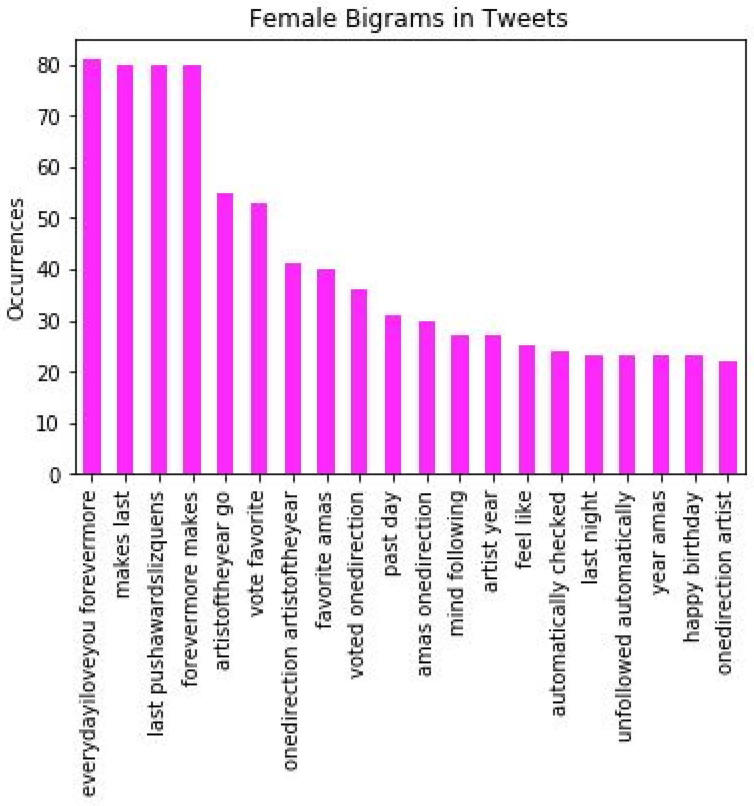


Figure 12: Female Words in Tweets

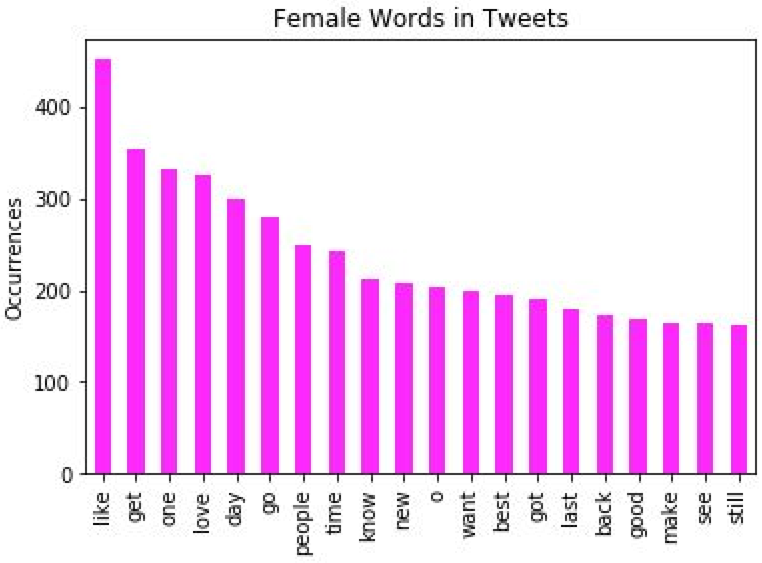


Figure 13: Male Bigrams in Descriptions

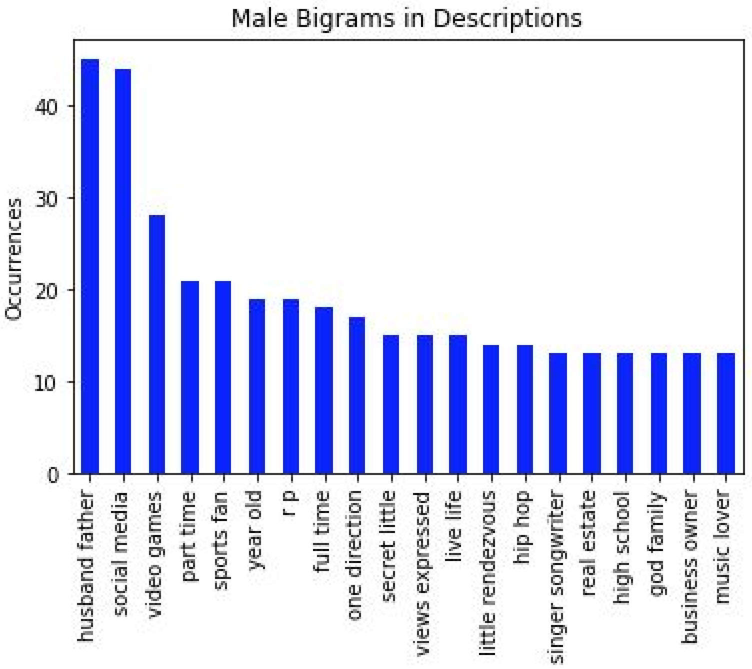


Figure 14: Male Words in Descriptions

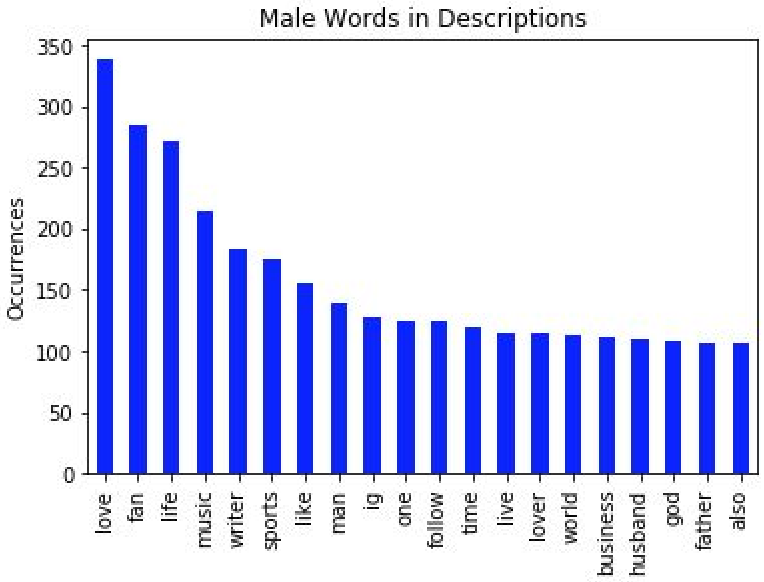


Figure 15: Male Bigrams Overall

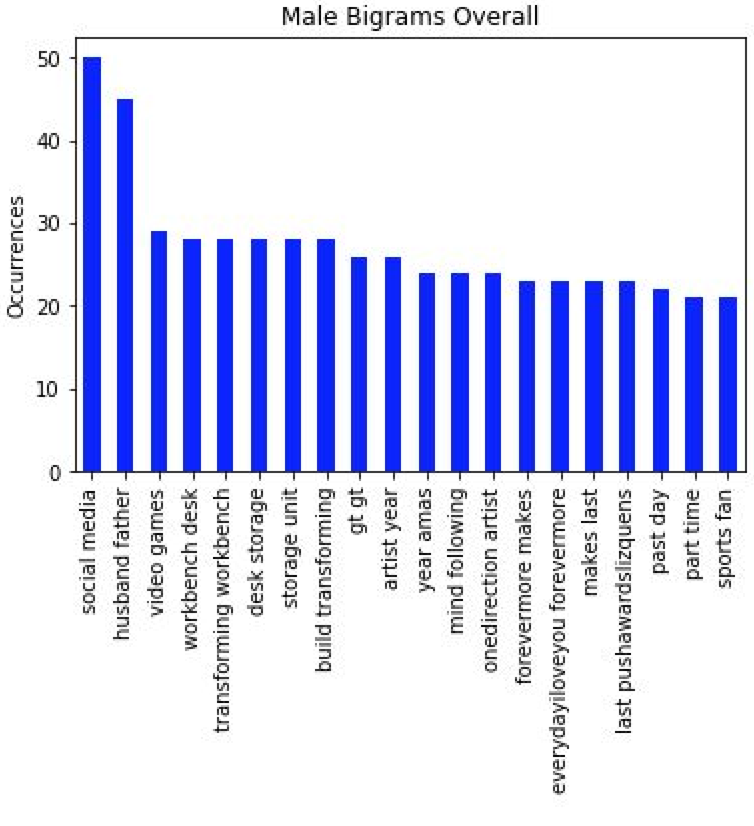


Figure 16: Male Words Overall

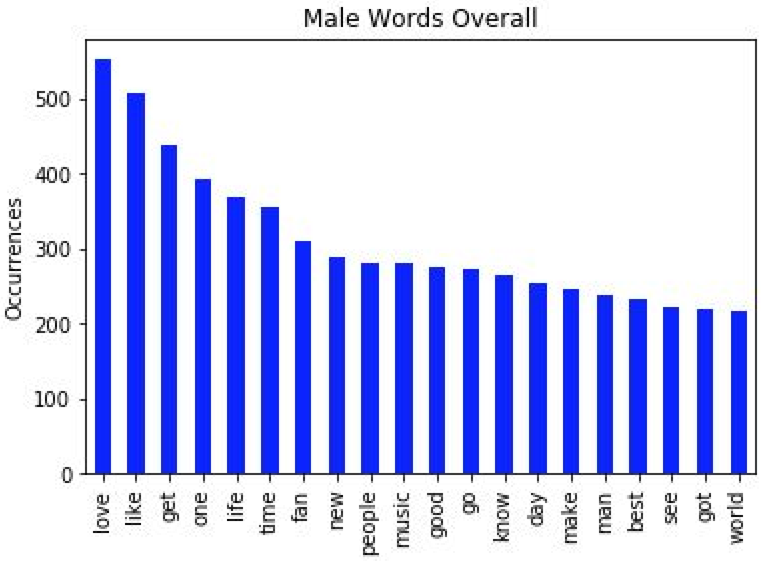


Figure 17: Male Bigrams in Tweets

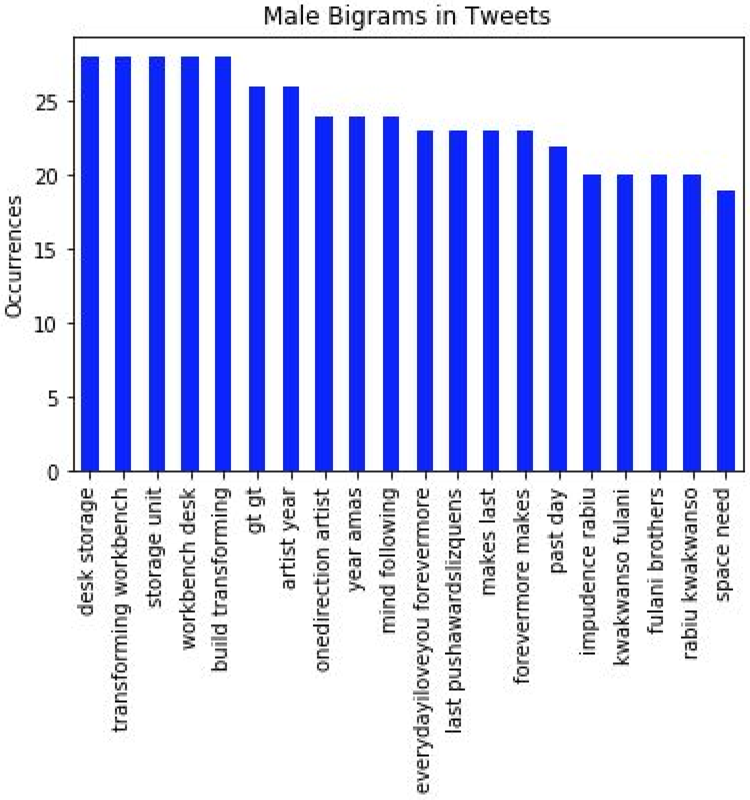
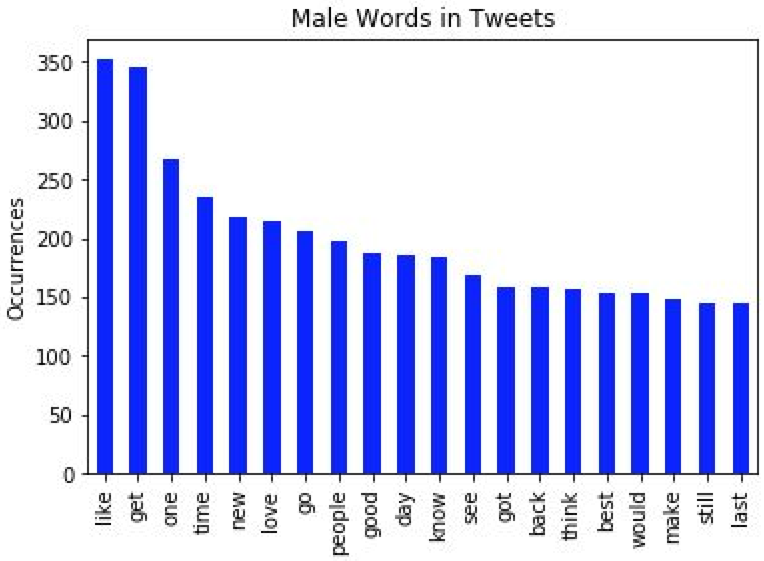


Figure 18: Male Words in Tweets


Male Words in Tweets
Occurrences
0
50
100
150
200
250
300
350
like
get
one
time
new
love
go
people
good
day
know
see
got
back
think
best
would
make
still
last