Gender Identification via Text Analysis

Data Mining Final Project - Sahil Shah, Moein Farokhnia, Negar Maleki, Syed Raza

## Introduction

There has been a sharp rise over the last several years in the prominence of gender detection and its implications as a social construct vs a scientific, biological phenomenon. The debate continues regarding the definition and significance of gender on an array of topics, and is at the root of many societal changes that are being driven by legislation across the globe.

The rise of this issue’s prominence has also caused a rise in the phenomenon known as “Catfishing” wherein people create fake virtual personas for various purposes.These purposes range from benign exploration of personal identity to malevolent sexual pursuit of minors. One form of catfishing is pretending to be the opposite gender on virtual platforms for various purposes.

This project explores the question of whether or not classical gender identity (male/female) can be predicted by the way people use words. There are many possible use cases for a predictive model that can accurately determine someone’s gender via text analysis. Use cases include, but are not limited to:

1. Identifying sexual predators on online platforms
2. Identifying individuals that are bad-actors or Cat fishing with a malicious social intent
3. Exploring the development of speech/writing through the lens of gender identity

To explore this topic our group created a novel text corpus to be used as the dataset for analysis. We did this by scraping over 1000 tweets from twitter and noted some key characteristics including emotion, type and gender . A TF-IDF score was assigned to all words in the corpus, and the output value was used as the primary prediction variable for five different prediction models.

Our results indicate that we can predict gender classification with an accuracy of ~60-65% with our dataset. There are several interesting differences in results from the various models deployed. It is important to note that this paper details the preliminary analysis conducted. There are several additional parameters for the emotion and category variables such as sarcasm and jokes that we could assign. New variables accounting for race and background could also be added in order to improve accuracy in further analyses.

The details of the data set, limitations, methodology, results, interpretations, possible improvements, and further questions and insights are discussed in detail in subsequent sections of this paper.

## 2. Data

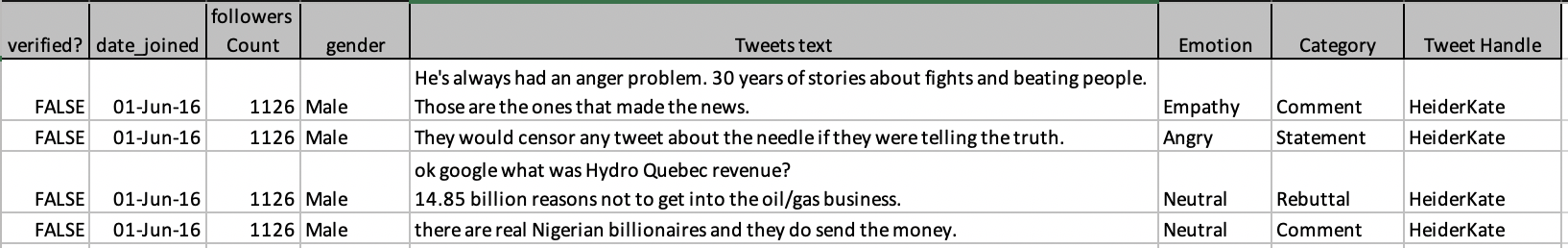
The dataset consists of 1,064 twitter posts. Each is labeled with the following attributes:

* Twitter handle (username)
* Verification status
* Date Joined
* Follower Count
* Gender (Male/Female)
* Tweet Text (Bag of Words)
* Emotion (happy, sad, angry, etc)
* Category (question, comment, rebuttal, statement)

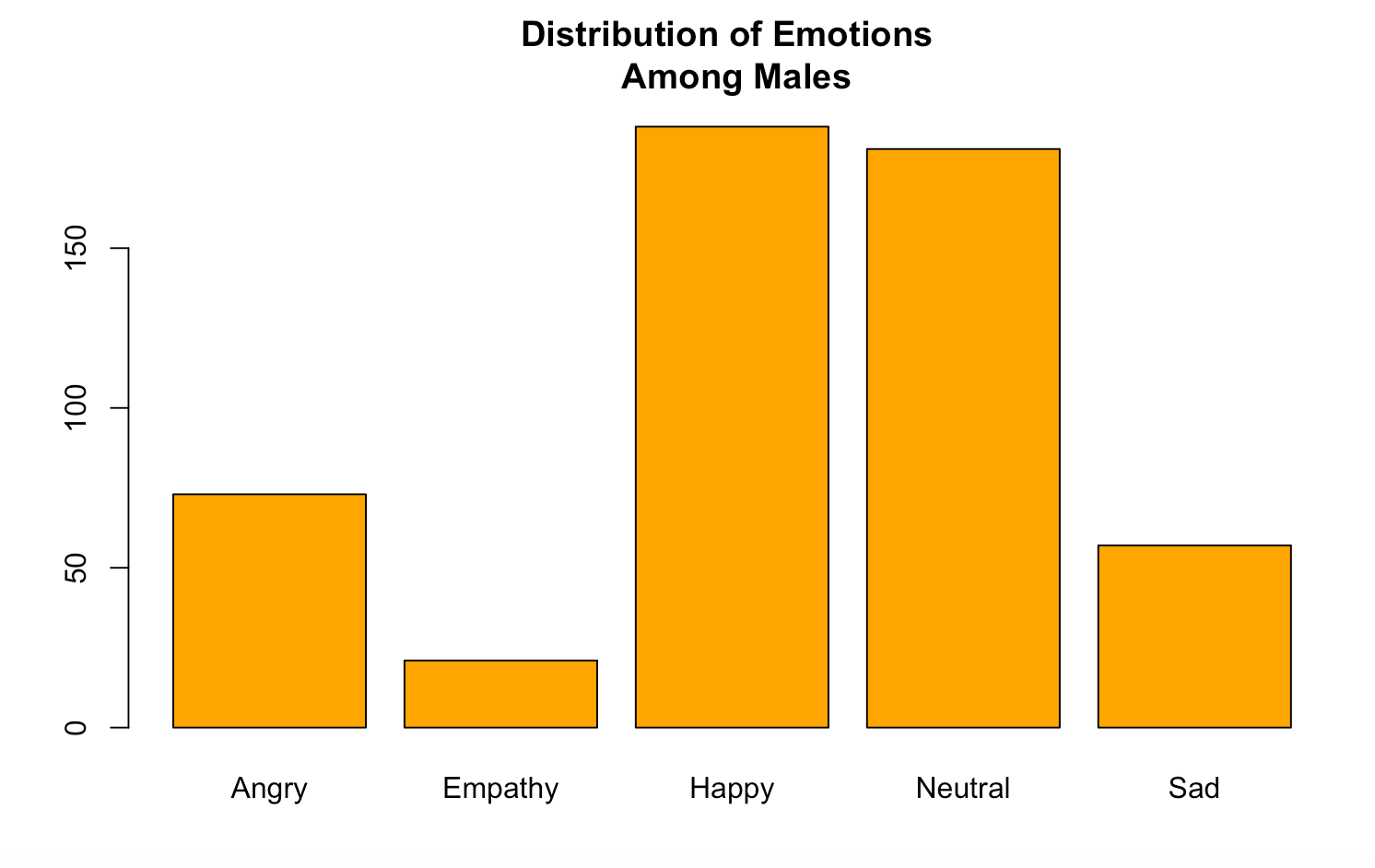
This dataset was assembled manually by scraping tweets from twitter, organizing them into a Microsoft Access DB, and inputting values for label variables. The target variable for prediction is gender, which is also manually labeled for each tweet. Between 10-20 tweets were scraped for each twitter user . Our data set has 543 observations for males, and 541 observations for females. The **primary objective of the manual data scrape was to ensure validity of gender assignment**.

Additional attributes such as emotion and category are intended to be used in further analyses to better capture key contextual cues that could be associated with gender. However, our analysis does incorporate emotion through the use of emojis as all emojis were converted into their text counterparts when creating the database.

A few rows of the data set are shown below



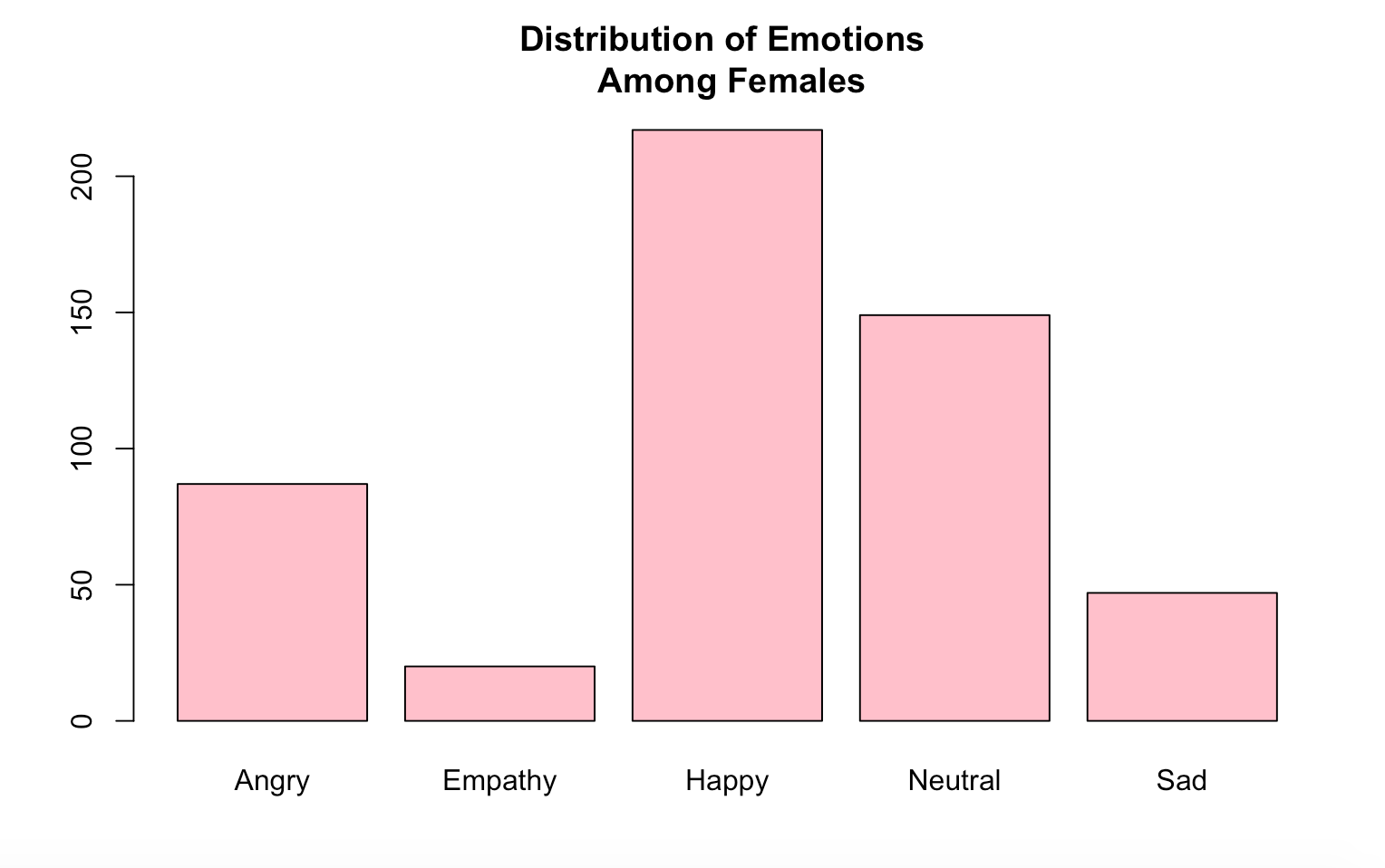
Distributions of emotions for males and females are shown in the following figures.



**> malesetemocount**

**Angry Empathy Happy Neutral Sad**

**73 21 188 181 57**



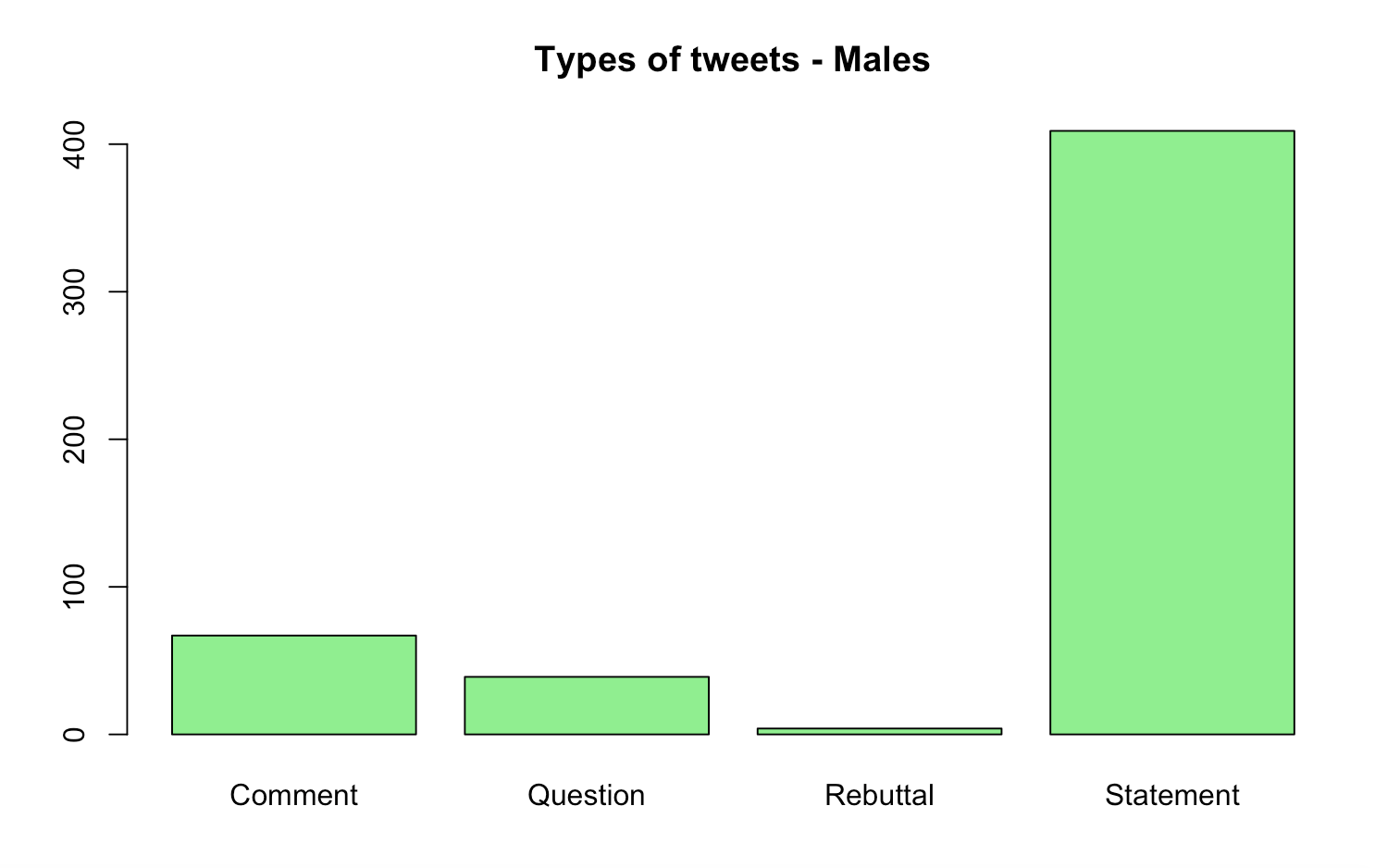
**> femalesetemocount**

**Angry Empathy Happy Neutral Sad**

**87 20 217 149 47**

There are some notable differences in emotions between the genders, namely a higher number of angry and happy tweets among females vs men. Men seem to post more “neutral” tweets than women.

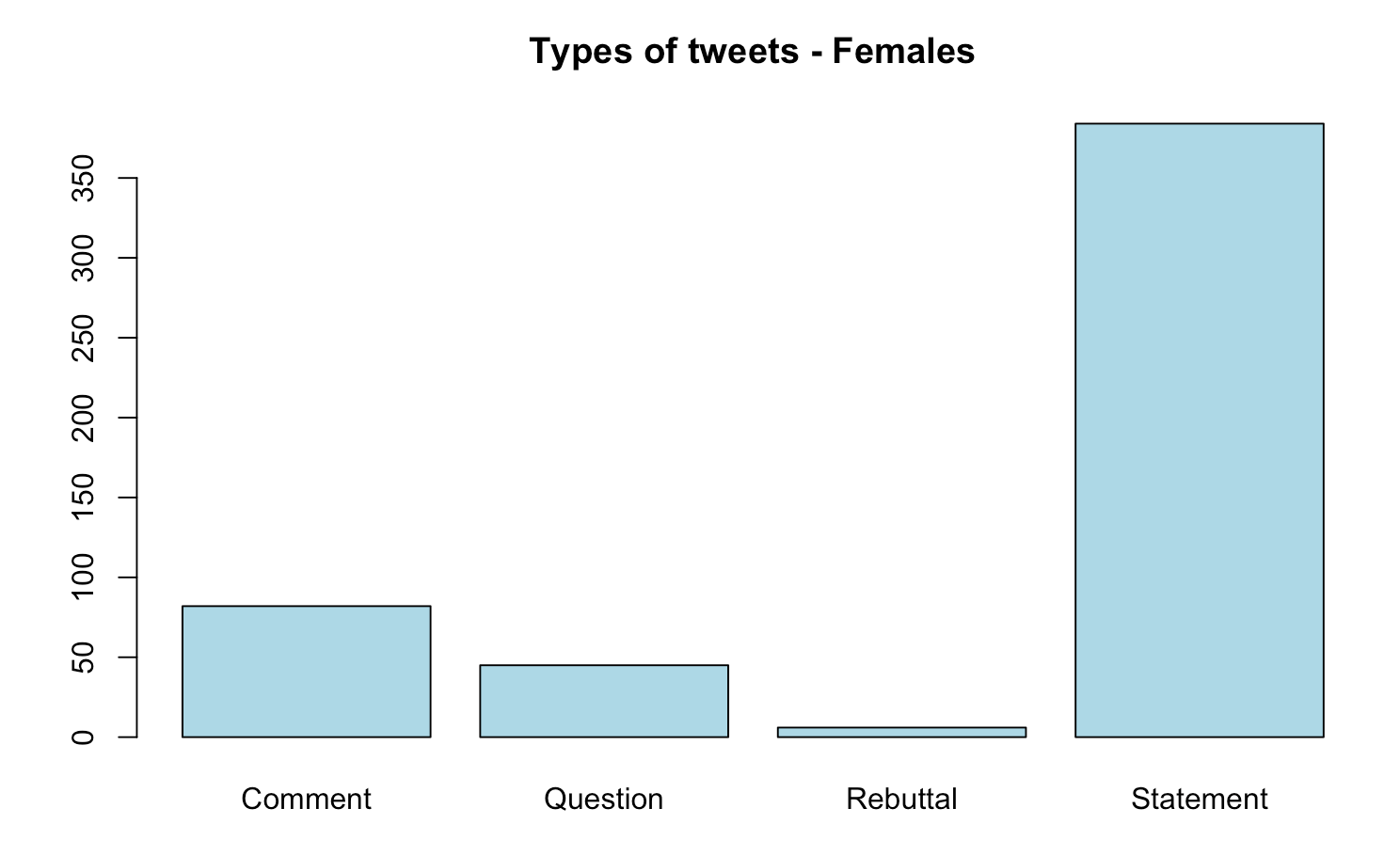
Distributions of categories (type of tweets) for men and women are shown in the following figures



**> malecategorycount**

**Comment Question Rebuttal Statement**

**67 39 4 409**



**> femalecategorycount**

**Comment Question Rebuttal Statement**

**82 45 6 384**

>70% of tweets for both genders were classified as statements. The trend of tweet category classification is similar for both genders - Statements > Comments > Questions > Rebuttal

**Limitations**

There are several important points to mention regarding the limitation of the current dataset and analysis.

1. **Manual data scrape bias**

Since we manually created this dataset by scraping tweets from twitter, the dataset is prone to inherent bias based on individual group members’ choices of whose tweets to scrape. For further analysis we intend to use the Twitter API to randomly scrape tweets and then identify the gender of each user. The purpose of our current scrape was to ensure the validity of gender assignment, and for that purpose choosing people who we can confirm associate to one gender or the other.

1. **Lack of grammar / lexicon analysis**

Our analysis relies upon the TF-IDF score for words within the corpus which have been stemmed.

TF-IDF based modeling predicts on a singular word level, and does not take into account prior words, or strings of words. As such, it omits the grammatical context, which is important in a comprehensive analysis.

Analyzing the stemmed corpus would not capture grammatical differences between genders, i.e. the way genders use gerunds and other word transforms, as well as the use of punctuation such as exclamation points.

1. **Subjective labeling**

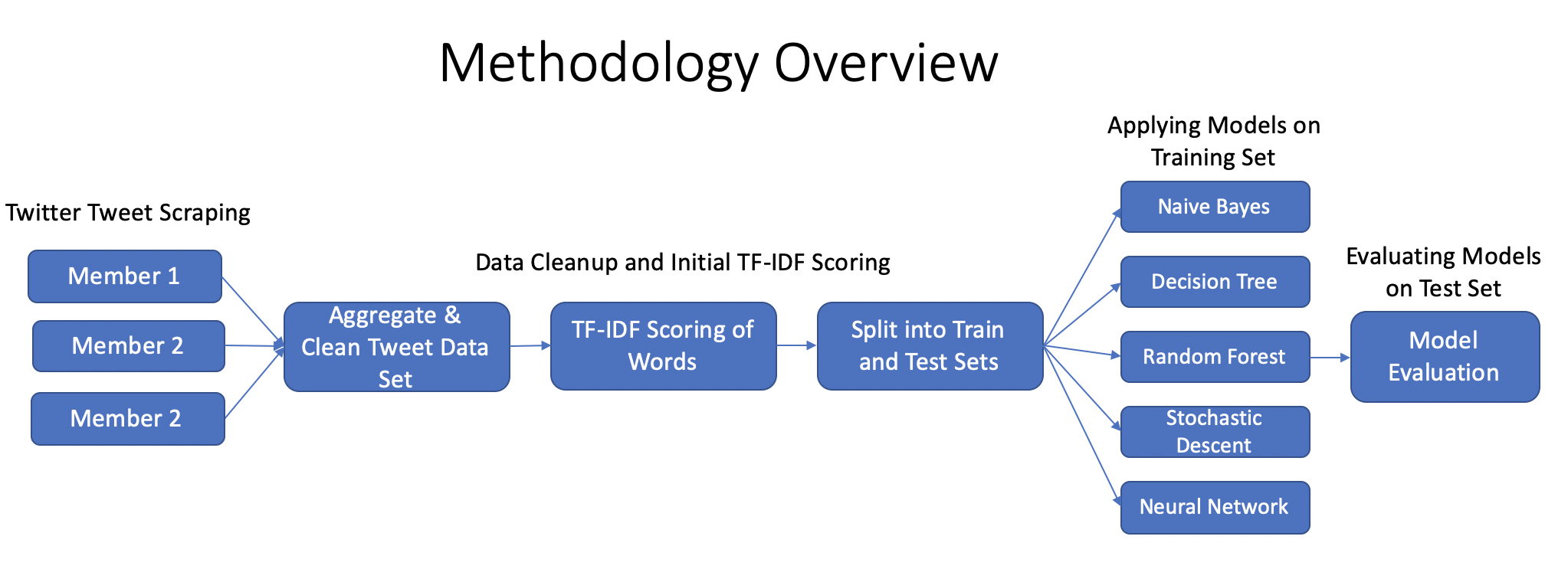
The labeling of the emotion and category variables are done based on the group member’s subjective choice. For a more objective labeling we would need to formally define the verbal metrics for each individual emotion. As a consequence the data set may be artificially skewed towards one emotion or another.

1. **Inability to ascertain particular words/phrases used for Decision Tree/Random Forest splits**

Since we implemented SVM on the total data set to reduce features, we are unable to ascertain which words/phrases svm produced in the final 676 feature set. As such we cannot dig much deeper than the TF-IDF score on individual words to show which words were used for node splitting.

## 3. Methodology & Results

3a. OVERVIEW



The overall process consisted of steps as shown in the figure above. They were

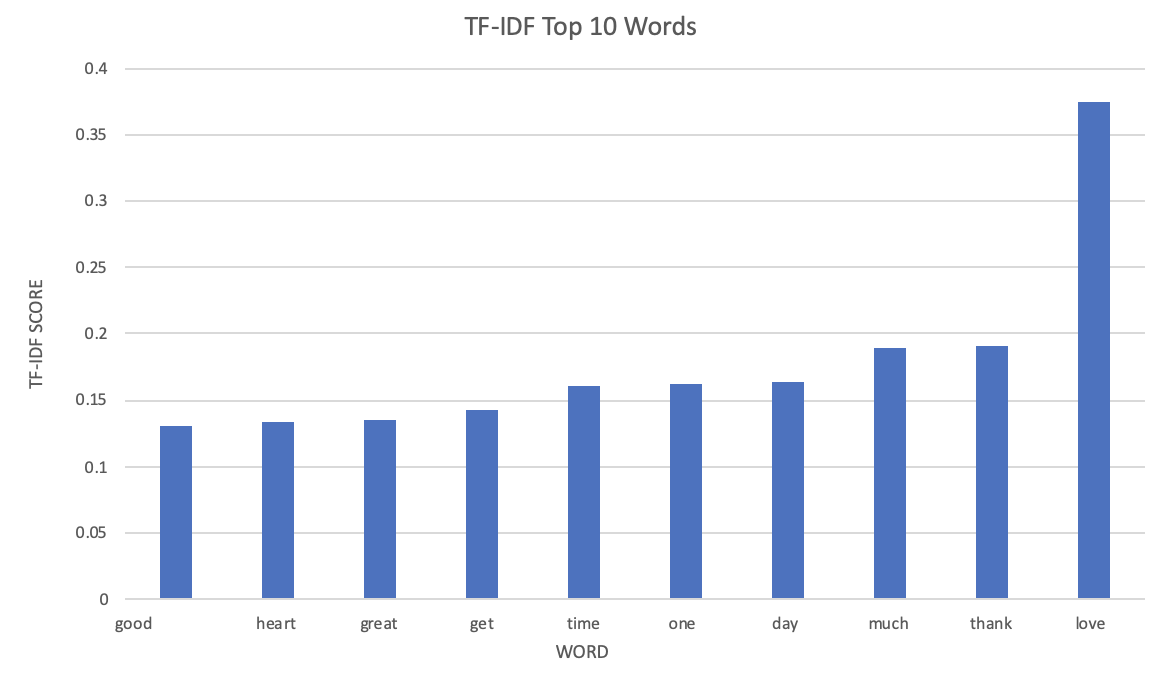
1. Data collection, aggregation, cleaning
2. Splitting into train and test sets
3. Count Vectorization of corpus
4. TF-IDF Scoring of words
5. SVD
6. Training the models
7. Evaluating models on test set

3b. PREPROCESSING, TRAIN, TEST, EVALUATION

Data preprocessing efforts were primarily sequestered to minimizing missing values for gender, removing stop words, and stemming the corpus.

For the purposes of this preliminary analysis we split the complete dataset into **train and test sets of 75% and 25% respectively**. The **original dataset had ~8000 words,** each considered its own feature. We applied a **CountVectorizer** on all words in both the train and test sets.

We then applied the **TF-IDF transform on both the train and test sets** and thendeployed five different machine learning models using the TF-IDF scoring as the primary predictive variable. The Top 10 TF-IDF Scoring words are shown in the graph below.



After applying TF-IDF **we used Singular Value Decomposition (SVD)** **which reduced the total number of features to 676. These 676 features are not necessarily singular words. They could be highly correlated strings of words that appear often together.**

We then trained five different machine learning models, and evaluated their performance on the test set. The five models used and corresponding accuracy on the test set are shown in the table below

**Models Deployed and Corresponding Accuracy**

| **Model** | **Male Acc** | **Female Acc** |
| --- | --- | --- |
| Naïve Bayes | 68% | 63% |
| Random Forest | 56% | 63% |
| Stochastic Gradient Descent Classifier | 63% | 63% |
| Decision Tree | 54% | 53% |
| Neural Network | 62% | 62% |

NAIVE BAYES:

The Naive Bayes Classification model produced the highest overall precision for male gender prediction at 68% and tied for highest precision for female gender prediction at 63%. The Naive Bayes assumption here is that the probability of each count vectorized feature is independent of all other features. This is interesting as it indicates that singular words/phrases can be used to predict gender, meaning that each gender seems to have a proclivity to use certain words/phrases, or combination of words (though not necessarily in a particular order).

Although the overall precision is highest for Naive bayes, the confusion matrix of the output indicates that we have a high False Negative rate

Model 1: Naive Bayes

precision recall f1-score support

Male 0.68 0.58 0.62 130

Female 0.63 0.73 0.68 129

Accuracy 0.65 259

Macro avg 0.66 0.65 0.65 259

Weighted avg 0.66 0.65 0.65 259

| Actual/Predicted | Female (1) | Male (0) |
| --- | --- | --- |
| Female (1) | 75 (TP) | 55 (FP) |
| Male (0) | 35 (FN) | 94 (TN) |

DECISION TREE & RANDOM FOREST

Decision tree classification model resulted in the lowest accuracy for both male and female gender. The decision tree created 106 nodes and leaves. However by changing the split on each node we can get better results as random forest models yield better results. Random forest results in 10 percent increase in accuracy for Female gender prediction however it does almost the same as the decision for Male gender.

Model 2: Decision Tree

precision recall f1-score support

Male 0.54 0.51 0.52 130

Female 0.53 0.57 0.55 129

Accuracy 0.54 259

Macro avg 0.54 0.54 0.54 259

Weighted avg 0.54 0.54 0.54 259

confusion matrix:

| Actual/Predicted | Female (1) | Male (0) |
| --- | --- | --- |
| Female (1) | 66 (TP) | 64 (FP) |
| Male (0) | 56 (FN) | 73 (TN) |

Model 3: Random Forest

precision recall f1-score support

Male 0.56 0.78 0.65 130

Female 0.63 0.39 0.48 120

Accuracy 0.58 259

Macro avg 0.60 0.58 0.57 259

Weighted avg 0.60 0.58 0.57 259

confusion matrix:

| Actual/Predicted | Female (1) | Male (0) |
| --- | --- | --- |
| Female (1) | 101 (TP) | 29 (FP) |
| Male (0) | 79 (FN) | 50 (TN) |

STOCHASTIC GRADIENT DESCENT

The SGD model produced results similar to the Neural Net which is interesting as it employs a similar approach to analysis, parameterizing an objective function and then working to optimize the model in the opposite direction of the objective function gradient. It does this iterative process for small batches of data, and in this way differs significantly from the Neural Net which iterates based on the output on the entire dataset.

In our case we used default parameter values and tried two loss functions. We tried both hinge (linear SVM) and log (logistic regression) but found highest accuracy with hinge. Like Neural Nets, the logic of SGD is often hard to interpret in the case of specific analyses. The results of SGD using the hinge loss function with other default parameters are shown below.

Model 4: Stochastic Gradient Descent Classifier (Hinge)

precision recall f1-score support

Male 0.63 0.65 0.64 130

Female 0.63 0.62 0.63 129

accuracy 0.63 259

macro avg 0.63 0.63 0.63 259

weighted avg 0.63 0.63 0.63 259

confusion matrix:

| Actual/Predicted | Female (1) | Male (0) |
| --- | --- | --- |
| Female (1) | 84 (TP) | 46 (FP) |
| Male (0) | 49 (FN) | 80 (TN) |

NEURAL NETWORK

We used a neural network with 3 hidden layers. The hidden layers had 600, 400, and 200 nodes each. We kept the learning rate, tolerance and momentum at the default value of 0.001, 0.00001 and 0.9 respectively. We adjusted the number of maximum iterations to 1000.

Neural Nets are often highly accurate but very hard to interpret. Although the accuracy of the Neural Net was not as high as some of the other models, it had the most consistent results for prediction between males and females as indicated by the confusion matrix. The results of the Neural net model are shown below.

Model 5: Neural Network

Precision recall f1-score support

Male 0.62 0.62 0.62 130

Female 0.62 0.62 0.62 129

Accuracy 0.62 259

Macro avg 0.62 0.62 0.62 259

Weighted avg 0.62 0.62 0.62 259

confusion matrix:

| Actual/Predicted | Female (1) | Male (0) |
| --- | --- | --- |
| Female (1) | 81 (TP) | 49 (FP) |
| Male (0) | 49 (FN) | 80 (TN) |

## 4. Related Work

-Who else has done related work here? What have they done and what were the findings? How(https://aclanthology.org/C14-1184.pdf)

Dong Nguyen has a study where they tried to predict not only gender but age as well. The goal was to study sociolinguistics, observing and quantifying the relationship of gender and age with language. Eckart, who is referenced in the publication, “gender and age are shaped by societal context..” suggests that they should be considered fluid variables that can change with culture and society.

The researchers collected data in an innovative way by introducing a game that would allow users to guess the gender of a tweet given to them. They would later use this prediction obtained and compare it with their own model to test its accuracy. The team here used a logistic regression model to predict the gender of the user. The results showed that users who guessed gender on 10 tweets in the game had an average accuracy of 71%, whereas their model had an average accuracy of 69%. The team hypothesized that the reason the results for the prediction model were less than that of the humans was because the dataset had only 20-40 tweets per user, suggesting that a larger dataset would likely train the model for higher accuracy.

Our methodology for the current analysis does not include a human equivalence component. This is likely one of the major next steps we would take in assessing whether or not our models can be considered accurate. However, comparing our Naive-bayes model to the human trial in the Nguyen study shows similar results to their logistic regression model. Though we did not use logistic regression models explicitly, we did use an SGD model with a logistic regression loss function. However the results were not as accurate as the SGD hinge, and thus were not included in this paper.

There are multiple other publications using text analysis on tweets to determine sentimentality. We had initially planned to use a twitter tweet set that was used for sentiment analysis but then due to lack of gender classification, we chose to build our own.

## 5. CONCLUSIONS AND NEXT STEPS

**SUMMARY**

In this project we created models to predict users’ gender based on the context of their tweets. The five different models produced different results, which are detailed in the methodology and results section above. The range of accuracy varied from 53% - 68%. One of the objectives of this endeavor was to compare different models.

The model with the highest accuracy was Naive Bayes at 68% for males and 63% for females. However there was a relatively high proportion of false negatives (true males identified as females) and false positives (true females identified as males). The higher accuracy of Naive-Bayes is particularly interesting as it suggests that single words/phrases may be strong predictors of gender. Furthermore, though overall accuracy was highest with NB, the misclassification rates were considerably different for males and females.

The model with the most consistent results was the Neural Net, whose confusion matrix was nearly identical for males and females. This may suggest that if we modify the Neural Net and produce higher accuracy, the results would be consistent for both males and females.

**Plans for further analysis:**

This preliminary analysis opens the door to several paths for deeper analysis and understanding. They include more robust randomization of tweet scraping, context/lexicon analysis and analysis of punctuation.

Unlike traditional numeric datasets, language presents several unique challenges, especially with regard to phrasing and grammar. The overarching objective is to facilitate contextual analysis via the implementation of ML algorithms to find patterns in language nuances that differ between genders.

Some of the next steps that we foresee to make this analysis more robust and applicable includes utilizing an LSTM model that can take into account prior word usage and thus analyze multiple word phrases. This would provide a better foundation for context analysis. Also, we intend to include certain stop words/punctuation in the analysis to find patterns that can reliably predict gender.

To reduce human selection bias, we plan to create a new dataset that utilizes the Twitter API to randomly scrape tweets and then assign gender to them. For more relevant accuracy evaluation we plan to have people predict gender on tweets and use the average prediction results as a benchmark for comparison.

**Practical use case application:**

In order to make the results viable for use cases like gender identification of potential online predators/bad actors we would need to improve the accuracy considerably. The cost function of misclassification here (incorrectly identifying gender misrepresentation) is likely very high, as it could result in misallocated criminal charges. As such we believe gender classification tools should be used in conjunction with other indicative measures to validate predictions.

The scope for accurate gender prediction is vast and highly desirable. It has many applications in real-world scenarios and could potentially have a significant impact on the way we understand gender itself and its implications as both a social and biological phenomenon.

## 6. Appendix

(Code fragments, pseudo-code, ML workflow diagrams etc

References:

# [Lara L.JonesLee H.WurmGregory A.NorvilleKate L.Mullins](https://www.sciencedirect.com/science/article/abs/pii/S0747563220300595#!), *Sex differences in emoji use, familiarity, and valence,* July 2020[https://www.sciencedirect.com /science/article/abs/pii/S0747563220300595](https://www.sciencedirect.com/science/article/abs/pii/S0747563220300595)

Nguyen, Dong. *Why Gender and Age Prediction from ... - Aclanthology.org*. https://aclanthology.org/C14-1184.pdf.

RESULTS:

Model 1: Naive Bayes

precision recall f1-score support

Male 0.68 0.58 0.62 130

Female 0.63 0.73 0.68 129

accuracy 0.65 259

macro avg 0.66 0.65 0.65 259

weighted avg 0.66 0.65 0.65 259

confusion matrix:

[[75 55]

[35 94]]

Model 2: Random Forest

precision recall f1-score support

Male 0.56 0.78 0.65 130

Female 0.63 0.39 0.48 129

accuracy 0.58 259

macro avg 0.60 0.58 0.57 259

weighted avg 0.60 0.58 0.57 259

confusion matrix:

[[101 29]

[ 79 50]]

Model 3: Stochastic Gradient Descent Classifier

precision recall f1-score support

Male 0.63 0.65 0.64 130

Female 0.63 0.62 0.63 129

accuracy 0.63 259

macro avg 0.63 0.63 0.63 259

weighted avg 0.63 0.63 0.63 259

confusion matrix:

[[84 46]

[49 80]]

Model 4: Decision Tree

precision recall f1-score support

Male 0.54 0.51 0.52 130

Female 0.53 0.57 0.55 129

accuracy 0.54 259

macro avg 0.54 0.54 0.54 259

weighted avg 0.54 0.54 0.54 259

confusion matrix:

[[66 64]

[56 73]]

Model 5: Neural Network

precision recall f1-score support

Male 0.62 0.62 0.62 130

Female 0.62 0.62 0.62 129

accuracy 0.62 259

macro avg 0.62 0.62 0.62 259

weighted avg 0.62 0.62 0.62 259

confusion matrix:

[[81 49]

[49 80]]

DECISION TREE

