Gender Classification in Modern Society

**Executive Summary**

*Background*

Social media has become increasingly popular over recent years providing an abundance of unstructured data to be mined and analyzed. Using this data can allow for a better understanding of society and thus improving social interactions. Twitter is a popular and free social forum that allows members to create microblogs, referred to as “tweets”.[1] The human mind is complex, and analysis of tweets can provide insight to various behavior patterns of individuals. Communication has changed vastly over the years due to the increase in technology, thus deeming it imperative to better the understanding of human interaction in modern times.

*Problem Statement*

There are endless benefits to data mining and analyzing social media, specifically “tweets”. Making sense of common words can allow for improved marketing, understanding the relationship between users and various topics, and improved interaction in society. It is important to keep the focus of analysis on social media specific to epistemology, rather than exploitation and purposes of greed.[2]

*Scope*

The scope of this project includes preprocessing and text classification using “tweets” collected from Twitter, including the user profile description and the user gender, and evaluation of the results of various models.

**Method**

*Data Source*

The data used for this project consists of unstructured data and was obtained from Kaggle.[3] The data was provided by Data for Everyone Library on CrowdFlower.

*Data Import/Cleaning*

The data was imported using Python’s pandas library and cleaned accordingly. The original shape of the dataset was 20,050 rows with 26 columns. Rows containing null values were completely removed as the data amount after removal was still sufficient. The gender variable contained 4 classes: male, female, unknown and brand. As the focus of this project is on content between male and female, observations that were unknown or brand were excluded (Figure 1). The final dataset to be processed forward had a shape of 12,991 rows and 3 columns. The columns used in the final analysis were gender, description, and text.

*Exploratory Data Analysis (EDA)*

The initial exploratory analysis was performed after cleaning the data, where much of the data required preprocessing. Cleaning the text was imperative in order to remove punctuation and various special characters that would disrupt further analysis. Regular expression functions were used to clean the text, including converting all words to lowercase.

Additionally, stop words as defined by the natural language processing toolkit library (nltk) were removed. Stop words are those that may appear frequently, but do not provide any value to the analysis. Words that were less frequency were also removed as these could be considered our outliers. Knowing that this text is tweets, there may be some extraneous words that will not provide value, either.[4] The top 20 words by male and female were each extracted after stop words were removed, and many shared words can be observed (Figure 2 and Figure 3).

Stemming and lemmatization are frequently used in text analysis to sort out words that are similar to one another. This process removes parts of the word, keeping just the base or root word. For example, loves and lovely would both become love. Porter stemmer library from nltk was utilized for stemming of tweets and description.

Figure 1: Percent of Tweets by Gender

Chart, pie chart

Description automatically generated

Figure 2: Bar Chart of Top Tweet Words by Males

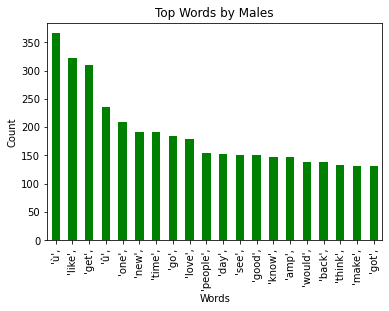
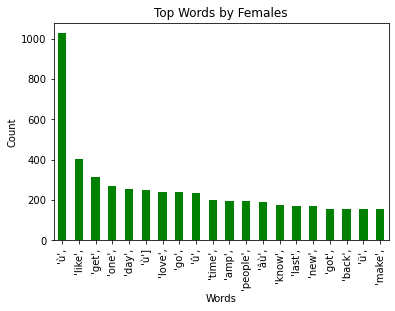
**

Figure 3: Bar Chart of Top Tweet Words by Females

**

*Feature Extraction & Vectorization*

Machine learning with text requires transforming the words in a document into numerical features, such that they can be fed to a model. This process is called feature extraction, more specifically, vectorization. Some methods that can be used for vectorization are frequency vectors which display a count of each word, one-hot encoding, a Boolean vector encoding method, and term frequency-inverse document frequency, which takes into account context by normalizing frequency of tokens within the document.[5] A count vector function was created using CountVectorizer library from sklearn where each input is preprocessed, tokenized at word level, and represented as a sparse matrix.[9] A function was created for this such that it could be used on Description and Tweets, separately. The gender variable was also converted to numerical, with one for female and zero for male.

*Model Deployment*

The data was split into a train and test dataset using the sklearn library. The following models were created to evaluate which approach gives the best accuracy; Naïve Bayes Gaussian, Naïve Bayes Multinomial, Logistic Regression, and XGBoost.

**Results**

The accuracy results from the four models explored are shown below. Interestingly, all models using gender and Twitter profile description had similar accuracy results, all within approximately 1% of each other. When the models were created using gender and a randomly selected tweet, the results varied a bit more. The XGBoost model had the highest accuracy results with 67.8%, and Logistic Regression had the lowest accuracy results with 55.6%.

Table 1: Modeling Results

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | |
| Gender and Tweets | Gender and Profile Description |
| Naïve Bayes Gaussian | 57.3% | 63.5% |
| Naïve Bayes Multinomial | 57.2% | 64.3% |
| Logistic Regression | 55.6% | 64.6% |
| XGBoost | 67.8% | 63.5% |

**Discussion**

*Assumptions*

An assumption made through the entirety of this project is that gender is binary, however, some may argue this is not the case. This project should be updated in the future when more data is available on a third binary gender option.

With Naïve Bayes classification, the assumption is made that each feature is independent and equal. This means that that none of the features depend on another, and that each feature used is equally important, and the attributes contribute equally to the outcome. It is noted though, that the assumptions may not be correct in real-world situations, but still work well in practice. Gaussian is a type of Naïve Bayes classification, where continuous values are assumed to have a Gaussian distribution, also known as a normal distribution appearing visually as a bell-shaped curve. [10]

Multinomial Naïve Bayes is another type of Naïve Bayes classification used for modeling with similar assumptions, but with the difference being that feature vectors are represented as frequencies. Naïve Bayes is commonly used and proven successful with document classification. [10]

Logistic regression is a type of classification that is used to predict a binary target variable. Binary classification is used when the outcome is of two options, binary, in this project, male or female. Logistic regression was used, rather than linear, as the purpose is predicting the most probable categorical result when given a tweet or profile description.[11] The assumptions with logistic regression are that there is little to no multicollinearity, which means that there isn’t any correlation between the prediction and response variables. However, model accuracy is not impacted, rather model stability. Another assumption for logistic regression is that all observations are independent from one another. Lastly, there should be no major outliers. [13]

Lastly, XGBoost machine learning algorithm was used which is an optimized distribution gradient boosting library. Boosting generally improves accuracy by combining a set of week learned by repeatedly reweighing. This means that when predictions are incorrect, they are weighted lower, allowing correct predictions to be weighted higher.[12] Boosting improved the accuracy considerably when tweets were used to predict gender. Interestingly, when using profile description to predict gender, the accuracy results of all models was similar.

*Limitations/Challenges*

With any project that has a social focus, it is important to be aware of ethical considerations through the project. Many data science projects that exist, including this one, focus on a binary result of gender, male or female. However, this could be deemed an incorrect generalization in today’s culture as the assumption of only male and female reinforces the idea that gender is binary. There is a large lack of data on non-binary gender as it has not always been recognized as a third gender option. In fact, it wasn’t until 2020 that a third gender option was acknowledged legally. On a federal level within the US, in February of 2020, a legislation was introduced that would add the third classification option to US passport applicants, and this is expected to be implemented by the end of 2021.[6, 7]

*Next Steps*

As a nonbinary gender option is a recent implementation in the US, in the future, I plan to further explore gender classification with the addition of a nonbinary category.

**References**

[1] Contributor, T. T. (2015, December 25). What is twitter? - definition from whatis.com. WhatIs.com. Retrieved September 13, 2021, from <https://whatis.techtarget.com/definition/Twitter>.

[2] Felt, M. (n.d.). *Social media and the social Sciences: How researchers employ big data Analytics - Mylynn Felt, 2016*. SAGE Journals. Retrieved September 15, 2021, from <https://journals.sagepub.com/doi/full/10.1177/2053951716645828>.

[3] Eight, F. (2016, November 21). *Twitter user gender classification*. Kaggle. Retrieved September 19, 2021, from <https://www.kaggle.com/crowdflower/twitter-user-gender-classification>.

[4] rashida048. (2021, June 5). Exploratory Data Analysis of Text Data Including Visualization. Regenerative. Retrieved September 19, 2021, from <https://regenerativetoday.com/exploratory-data-analysis-of-text-data-including-visualization-and-sentiment-analysis/>.

[5] Bengfort, B., Bilbro, R., &amp; Ojeda, T. (2018). Applied text analysis with python: Enabling language-aware data products with machine learning. O'Reilly Media, Inc.

[6] Wikimedia Foundation. (2021, August 17). Legal recognition of non-binary gender. Wikipedia. Retrieved September 19, 2021, from <https://en.wikipedia.org/wiki/Legal_recognition_of_non-binary_gender>.

[7] Cummings, W. (2017, June 28). When asked their sex, some are going with option 'x'. USA Today. Retrieved September 19, 2021, from <https://www.usatoday.com/story/news/2017/06/21/third-gender-option-non-binary/359260001/>.

[8] Sayyadiharikandeh, M., Ciampaglia, G. L., & Flammini, A. (n.d.). *Cross-domain Gender Detection in Twitter*. Retrieved September 9, 2021, from <https://glciampaglia.com/docs/papers/genderdetection_chasm.pdf>

[9] Jain, P. (2021, June 1). *Basics of countvectorizer*. Medium. Retrieved September 23, 2021, from <https://towardsdatascience.com/basics-of-countvectorizer-e26677900f9c>.

[10] *Naive Bayes classifiers*. GeeksforGeeks. (2020, May 15). Retrieved September 23, 2021, from <https://www.geeksforgeeks.org/naive-bayes-classifiers/>.

[11] Kvijayan, V. (2019, September 23). *Classification using logistic regression*. Medium. Retrieved September 25, 2021, from <https://medium.com/@vishnu0205/classification-using-logistic-regression-f14434d8545f>.

[12] *XGboost Python sklearn regression Classifier tutorial with code examples*. DataCamp Community. (n.d.). Retrieved September 25, 2021, from <https://www.datacamp.com/community/tutorials/xgboost-in-python>.

[13] Tung.M.Phung. (2020, July 31). *Assumptions of logistic regression*. Tung M Phung's Blog. Retrieved September 25, 2021, from <https://tungmphung.com/assumptions-of-logistic-regression/#:~:text=%20Assumptions%20of%20Logistic%20Regression%20%201%20Little,Logistic%20Regression%20is%20not%20famous%20for...%20More%20>.

**Appendix**

Below is a dataframe of the dataset after cleaning, with the far right two columns containing the clean data.

Graphical user interface, text, application

Description automatically generated

Below is an image of the “stop” words that were removed.

Text, letter

Description automatically generated