April 17, 2020

Twitter Gender Classification

Predict user gender based on Twitter profile information

Name: Shengnan(Elenore) Duan
Unique Name: elenore
UMID: 50457744

Table of Contents

1. Project overview

I found this Twitter User Gender Classification dataset online, and I think it would be interesting to predict user gender by multiple features. This dataset has many interesting features including both numerical data and categorical data, such as gender confidence and the sidebar color choices. The gender labels are also interesting, that they are not only just common genders (male and female), but also brand/organization and unknown. Therefore, by analyzing the data, I hope to answer the following questions:

- Q1: What are the words in tweets that strongly distinguish male or female gender?
- Q2: What are the important features to predict gender label?
- Q3: How well do stylistic features (color features) to predict gender label?
- Q4: How well do text features to predict gender label?

2. Data source

The link I downloaded the dataset is https://www.kaggle.com/crowdflower/twitter-user-gender-classification, it's originally provided by the Data For Everyone Library on Crowdflower.

2.1. Data description

Please refer to my source code file for full data description. The dataset is 8-megabyte big, containing 20050 rows and 26 columns. Each row with a user's name, a random tweet of this user, account profile and image, location, and even link and sidebar color etc. The dataset can be found in the link above or the zip file I submit. The label I want to predict is *gender* variable, there are 4 values of gender: male, female, brand and unknown.

Below are some important columns and corresponding data types:

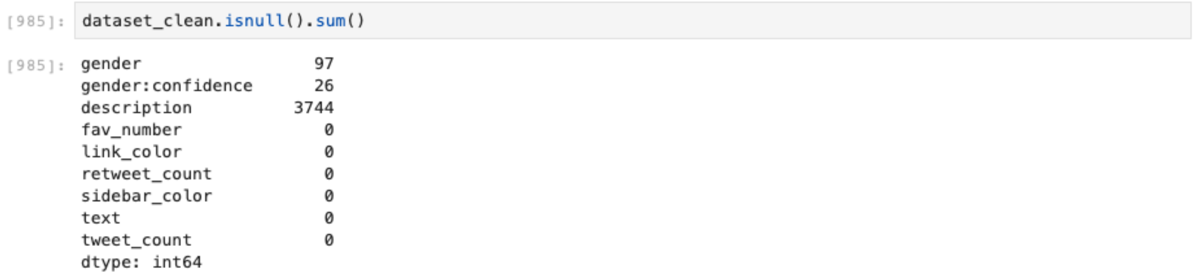
gender	Category
gender:confidence	Numeric
description	Category
fav_number	Numeric
link_color	Category
retweet_count	Numeric
sidebar_color	Category
text	Category
tweet_count	Numeric

2.2. Initial data exploration and cleaning

Firstly, I imported all the necessary packages: pandas and numpy for data manipulation in general, NLTK, Counter and Wordcloud for text data processing, matplotlib and seaborn for data visualization, sklearn for machine learning, and etc.

There are 26 columns in the dataset, but not all of them are useful. For example, _golden has 100% 'false' value, _unit_state has 100% 'finalized' value and _trusted_judgments have 100% '3' value (it's determinate by the _golden variable). I also dropped the date columns since they are not easy to process, and I don't think they are important features for gender prediction, so as columns like name, tweet_id, tweet_location and user_timezone. The ‘profileimage’ is not analyzable because it’s a link of profile picture.

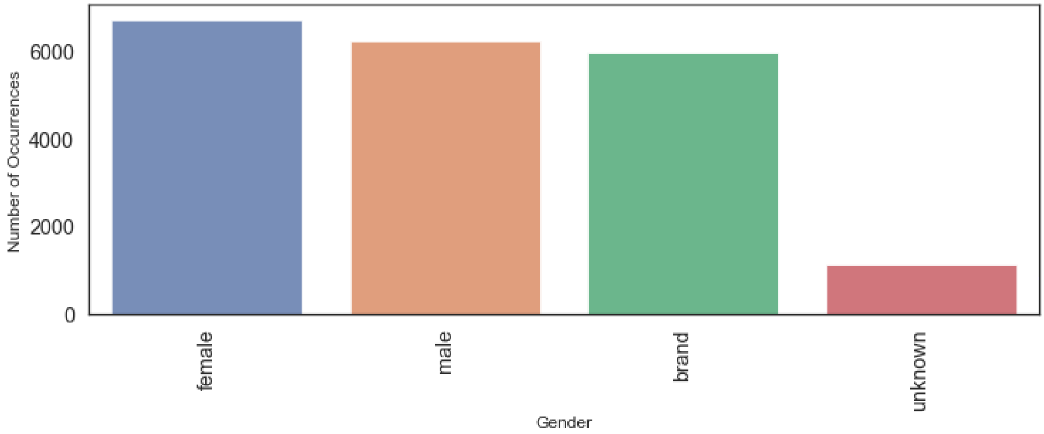
After the initial cleaning, I started to handle the missing value. I firstly checked how many n/a values in each column. Then I decide to drop n/a values for *gender* and *gender:confidence*, since they are not much compared with the total observations. But the n/a values for *description* column is a lot, and that’s reasonable because some users don’t put description in their profile, and that might be a gender preference, so I will process this column later not by dropping n/a.



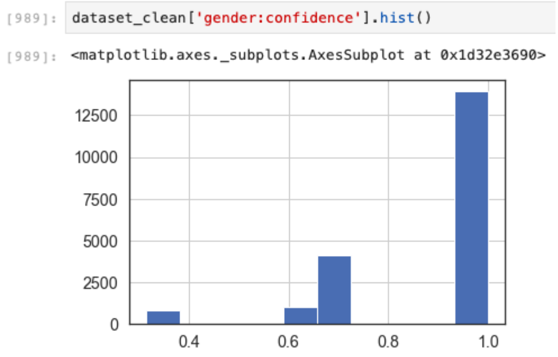
```
[985]: dataset_clean.isnull().sum()
```

```
[985]: gender                 97
       gender:confidence      26
       description          3744
       fav_number              0
       link_color              0
       retweet_count           0
       sidebar_color           0
       text                    0
       tweet_count             0
       dtype: int64
```

I checked the unique values for each column and decide to visualize the gender distribution. Male, female and brand users are quite evenly distributed, although unknown-gender uses are not much, I’ll keep it in the predictive labels, because it could represent the users who has difficulty about their gender identities.

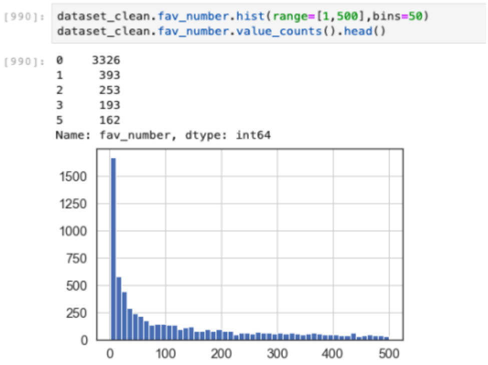


I also visualized the *gender:confidence* variable, and it can be seen that most people are quite confident about their gender, whereas not small amount (about 5000) of users are hesitating. As for *fav_number* column, 3326 users has 0 favorite tweet, accounting for the largest group, so when I visualize it, I choose the range starting from 1.



```
[989]: dataset_clean['gender:confidence'].hist()
```

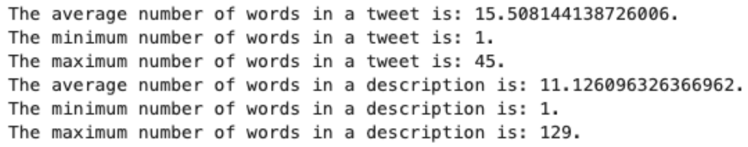
```
[989]: <matplotlib.axes._subplots.AxesSubplot at 0x1d32e3690>
```



```
[990]: dataset_clean.fav_number.hist(range=[1,500],bins=50)
       dataset_clean.fav_number.value_counts().head()
```

```
[990]: 0    3326
       1     393
       2     253
       3     193
       5     162
       Name: fav_number, dtype: int64
```

Next, I have a look at the text data. There are 2 text columns in the dataset: *description* (the profile description) and *text* (a random tweet). Instead of dropping the missing value, I fill the n/a value for *description* by a special word 'no_description'. The general overview of text data is as below. Lastly, I think the length of the text data could be an important feature, so I added 2 new columns as *text_length* and *description_length.* By then, there is no n/a value in the dataset, and the data is cleaned for analyzing in the next section.



```
The average number of words in a tweet is: 15.508144138726006.
The minimum number of words in a tweet is: 1.
The maximum number of words in a tweet is: 45.
The average number of words in a description is: 11.126096326366962.
The minimum number of words in a description is: 1.
The maximum number of words in a description is: 129.
```

3. Methods

For each question, the large part of preparation for manipulating data and handling missing/noise data is finished in the section 2.2.

3.1. Q1: What are the words in tweets that strongly distinguish male or female gender?

<u>Data cleaning</u>
The descriptions and tweets can be very messy text data, since users could use punctuation, special characters and non-English words. To clean the text, firstly, I use beautifulsoup and re packages to remove urls. Secondly, I removed the stop words and some customized punctuations, and filter away all non-English words. Next step is tokenization. By calling word_tokenize function from NLTK package, I added 2 new columns: *text_token* and *description_token*.

<u>Data analyzing</u>
I create the set of all token tweets (*text_token)* of female and male gender and transform them into string format. My plan is to use wordcloud to visualize the frequency of words. To make the visualization fun, I use the female and male icon images to define the shape and adjust the attributes accordingly. From the results, I realize there are large amount of overlapping words, and I can't find the most distinguish words for 2 groups, so I removed the overlaps and get the final results in both visualization and string format.

3.2. Q2: What are the important features to predict gender label?

<u>Data cleaning</u>
To prepare the data for machine learning analysis, the core challenge here is to transform categorical data to numerical data. I don't want to use pd.get_dummies function here, since it will turn each category value to a new column, which will expand the dataset size largely and slow down the analyzing process. Therefore, I focus on transforming the data type manually. I firstly transformed the stylistic variables' to numbers by cat.code function. Then I called doc2vec to vectorize all text data (both descriptions and texts).

<u>Data analyzing</u>
The new embedding text columns transformed the original list of token words to a list of numerical tuples, which is not suitable for machine learning analysis yet. So, I created the columns for each tuple, and at the end, each text variable turns into 300 embeddings columns (300 description embeddings and 300 text embeddings). Same method applied to both training dataset and test

dataset. Lastly, I use random forest model to get the overall predication and features ranking (for text embeddings, I use the mean value for both variables).

3.3. Q3: How well do stylistic features (color features) to predict gender label?

Data cleaning
I extract the stylistic features: *link_color* and *sidebar_color* for both training and testing dataset.
Data analyzing
I use random forest model to get stylistic features predication accuracy and changed the 'fold' to see the changes of accuracy. Lastly, I will compare the actual and predicting values by scatter plots.

3.4. Q4: How well do text features to predict gender label?

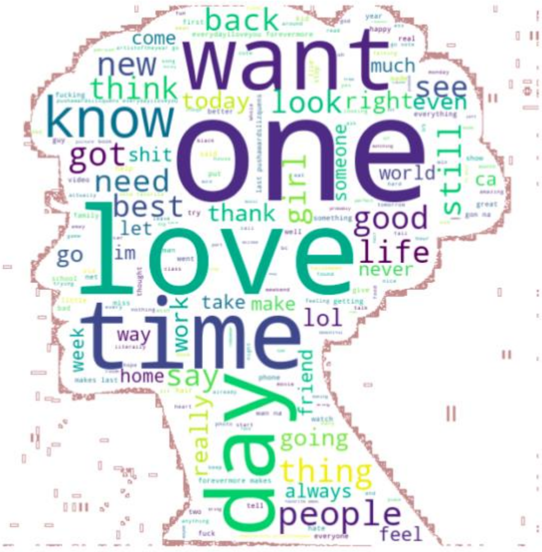
Data cleaning
I extract the text features: *600 embedding columns* for both training and testing dataset.

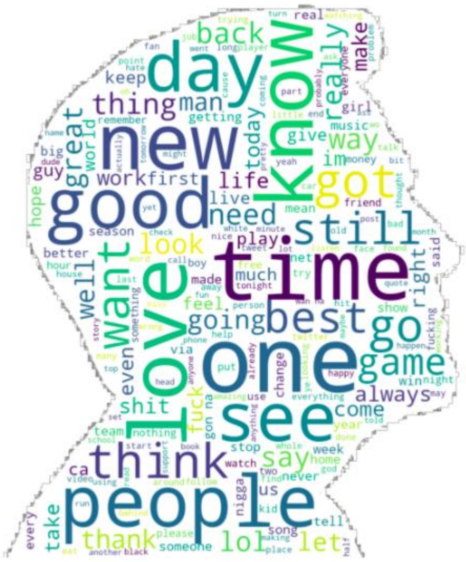
Data analyzing
I use random forest model to get stylistic features predication accuracy and changed the 'fold' to see the changes of accuracy. Lastly, I will compare the actual and predicting values by scatter plots.

4. Results

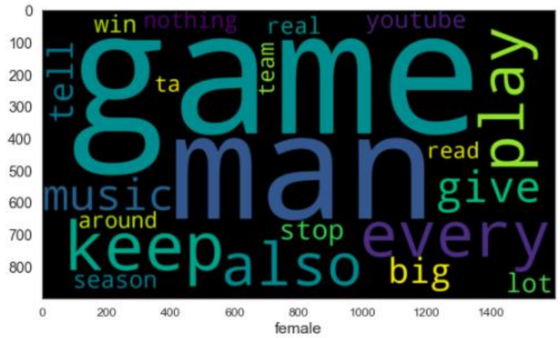
4.1 Q1: What are the words in tweets that strongly distinguish male or female gender?

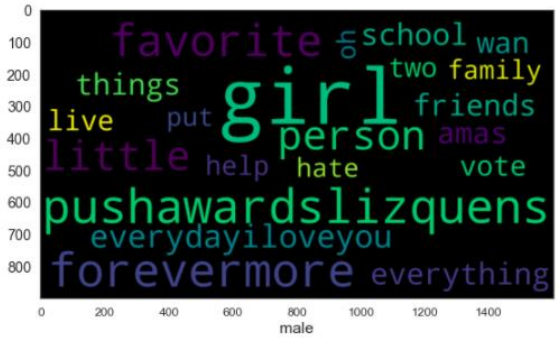
As mentioned in section 3.1, the initial results of most frequent words have some overlapping words of 2 genders. Even though, the frequency of those overlapping words can be different. For example, female twitter users use words 'love', 'one', 'want', 'day', more frequent than males, whereas males use words 'know', 'people', 'think', 'got', 'see' more often than females. In general, males have a wider range of common-used words than that of females.





Then I removed the overlapping common words, and distinguish words are as below:



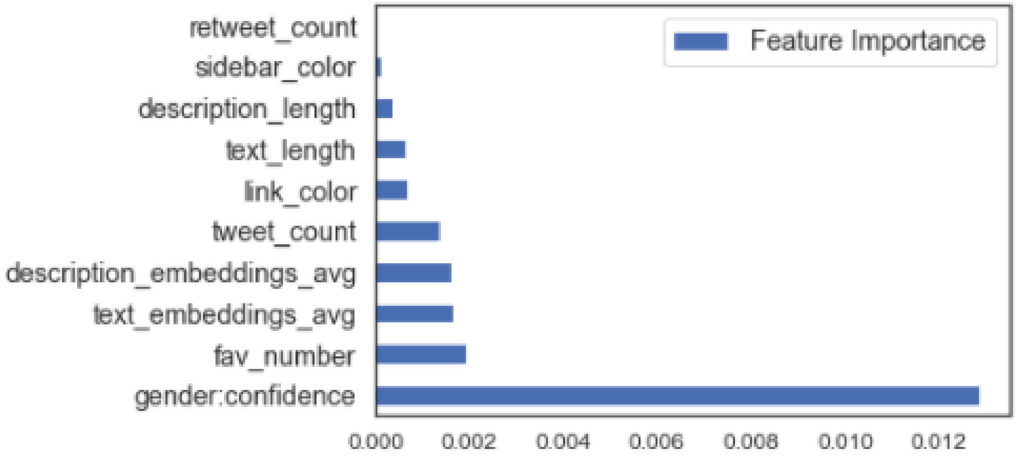


It seems females specifically more often talk about 'game', 'man', 'play', 'keep', 'music', 'every' while males are more likely to talk about 'girl', 'person', 'family', 'friends', and even more likely to use emotional hashtags like 'everydayiloveyou', 'forevermore' and 'pushawardslizquens' (not sure what this one means).

4.2 Q2: What are the important features to predict gender label?

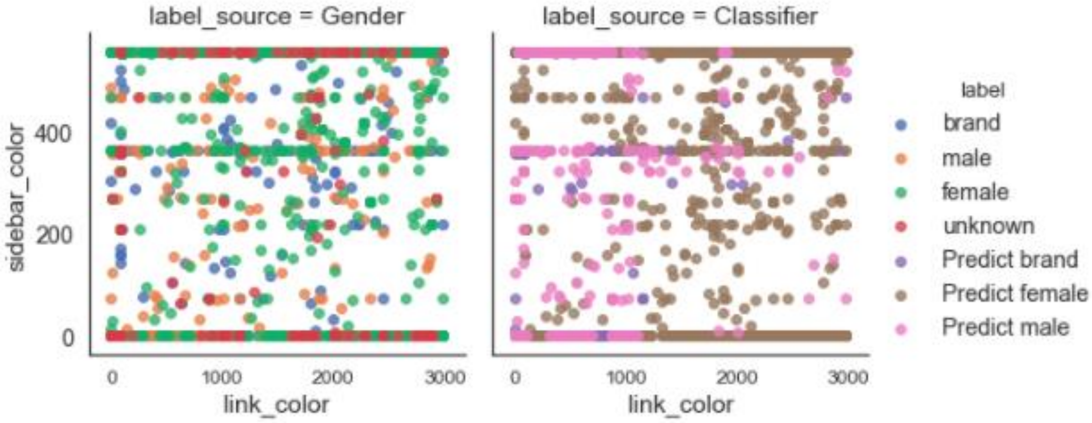
I include 10 features, both numerical and text data, to predict the gender label. The overall accuracy is around 55%, which seems not so impressive. However, to compare with the random guess accuracy rate 30%, the training process is still worthy.

As for feature importance, since it's not reasonable to count each text embedding as a feature, I get the average values of the 2 text features and labeled them as description_embeddings_avg and text_embeddings_avg. However, this might not be the best solution and can't show the true importance of text feature. Overall, according to the rank, *gender:confidence* is the most important feature.



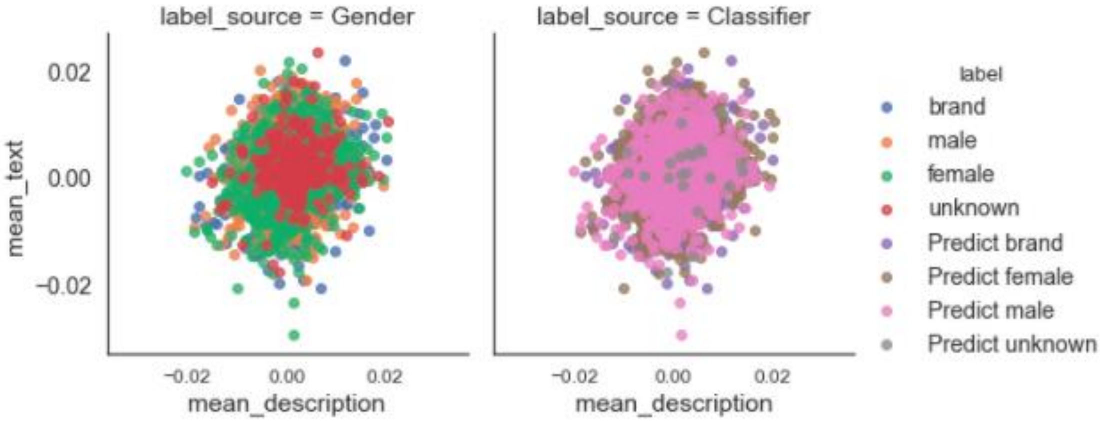
4.3 Q3: How well do stylistic features (color features) to predict gender label?

With the data cleaning process finished above, it's much easier to address Q3 and Q4. Among the 10 features, I extract features of stylistic and text group because they are user's intentional setting choices, whereas other features are more like objective numbers recording user's behaviors. My original assumption is female, male and brand users will have very different preference in colors, and the stylistic features could predict gender well. However, it turns out the accuracy is only 40% (still better than random guess).



4.4 Q4: How well do text features to predict gender label?

Similarly, I extract the text features and add 2 new columns as the average value of text embeddings for visualization later. I was assuming the tweets and description could be really different among genders, especially between private accounts and organization/brand accounts. It turns out it's right. The accuracy of text features is roughly 58%, even higher than the overall accuracy, and for sure higher than the random guess accuracy as well.



The accuracy is not ideal in this project and I hope I can improve my work after taking more machine learning classes. Also, I think if I change the feature selection/dataset, maybe I could get a better result.

References

1. https://www.kaggle.com/crowdflower/twitter-user-gender-classification

2. https://www.datacamp.com/community/tutorials/wordcloud-python

3. https://towardsdatascience.com/multi-class-text-classification-with-doc2vec-logistic-regression-9da9947b43f4