Gender Inference Based On Twitter Profiles

Francesco Ferrari
A53204320
Department of Computer Science
UC San Diego
fferrari@eng.ucsd.edu

Ayush Jasuja
A53103338
Department of Computer Science
UC San Diego
ayjasuja@eng.ucsd.edu

Manu Seth
A53219535
Department of Electrical and Computer Engineering
UC San Diego
mseth@eng.ucsd.edu

Ranti Dev Sharma
A53223197
Department of Computer Science
UC San Diego
rds004@eng.ucsd.edu

Abstract

The rapid growth of social networks has produced an unprecedented amount of user-generated data. Twitter, being one of the most accessible sources, is ideal for performing text analysis and determining profile demographics like gender, age and geographical location. In this study we seek to identify the gender of Twitter users based on their profile and one random tweet. Gender classification can serve multiple purposes. For example, commercial organizations can use gender classification for advertising. Law enforcement may use gender classification as part of legal investigations. Others may use gender information for social reasons. Our method makes use of users profile picture, profile description, link color, sidebar color apart from one tweet text to classify the profile as male, female or brand. A novel technique combining deep neural networks with classic text mining methods is employed to achieve an accuracy above 86%.

1 Similar Literature

The dataset is produced by Crowdflower, a data wrangling company, that aims to produce rich data for better analysis. In this case, contributors were asked to view the Twitter profiles and classify them as male, female and brand for 20,000 tweets. The model built on this data is then expected to produce similar results on other tweets, as it is built using a rich training data source.

[1] focuses only on colors used in the profiles at five different places and identify the band of colors that represent the male and female communities. [2] focuses mainly on tweets and uses N-gram character features to identify patterns of characters including acronyms and emoticons, used by the two communities. We found no literature which studies the combined effect of text, colors and images.

2 Dataset

We started our task by analyzing the details in the dataset that we had at our disposal. The dataset consisted of 20,000 rows collected from Twitter containing three different types of gender: male, female and brands (companies with an official Twitter account). Each row had a certain degree of confidence for the gender label. In order to make sure that we worked on the cleanest data possible we filtered every row that had a degree of confidence less than 1 and also every row that had unknown as gender. After filtering the dataset we had approximately 14,000 rows remaining with about 5,000 males and females and 4,000 brands.

The dataset was split further into a training set consisting of 80% of the rows and a validation set containing the remaining rows. The split between the training set and the validation set was consistent across all the different used models in order to make sure that we would get a validation accuracy over the same samples.

We had several features to work on which we will list below split in different types of categories.

Text-based features:

- username
- description of user profile
- one tweet selected randomly from the users tweets

Quantitative features:

- number of re-tweets
- link and sidebar color
- tweet location
- user's timezone

Images-based features:

- user's profile pictures

We approached each feature category in a different way since images and text require different types of pre-processing.

2.1 Building features from text

We started analyzing the text features by simply observing which words were most recurrent in each genders description. We did this but firstly separating the rows by gender, removing all the stop words and finally creating three wordclouds to plot them efficiently.

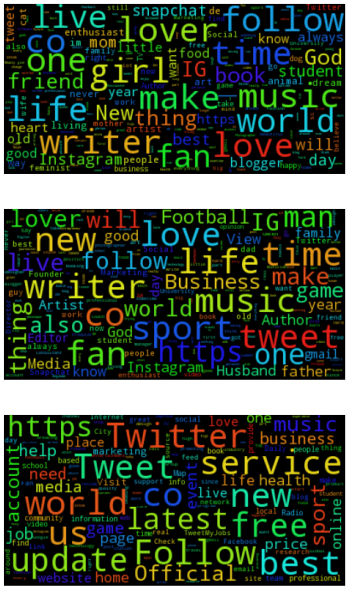

Fig 1: Wordcloud for females, males and brands based on their description.

From fig 1 we can see that, even though there are many terms that are shared across gender, there are also many others which are unique of each category such as girl, husband and official.

After confirming that the language does differ between gender we proceeded with a deeper analysis. We collected the top words appearing in the dataset, filter the ones that appear less than three times, and created a feature vector containing the top words and their frequency for each row. We then treated each row as a document used a python library called Gensim to apply Latent Dirichlet Allocation and observed the main topics that each gender talks about.

Once we obtained the average topic distribution for each gender we calculated the average topic distribution across the whole dataset. We then divided the topic distribution for each gender by the average one across the population and noted which topics per gender overindex or

underindex when compared to the general distribution. Below we show the top overindexing topics for each gender.

Word	$P(word\|topic)$
ig	0.1135
girl	0.0761
team	0.0430
friend	0.0324
watching	0.0220
smart	0.0207
kids	0.0193
mean	0.01808
needs	0.0176
dog	0.0175

Table 1: Top 10 words for top overindexing topic for females (they are 32% more likely to talk about it than average)

Word	$P(word\|topic)$
ig	0.1135
girl	0.0761
team	0.0430
friend	0.0324
watching	0.0220
smart	0.0207
kids	0.0193
mean	0.01808
needs	0.0176
dog	0.0175

Table 2: Top 10 words for top overindexing topic for males (they are 25% more likely to talk about it than average)

Word	$P(word\|topic)$
news	0.1722
people	0.0800
much	0.0364
latest	0.0308
think	0.0256
tv	0.0253
believe	0.0250
fans	0.0248
entertainment	0.0233
stories	0.0177

Table 3: Top 10 words for top overindexing topic for brands (they are 48% more likely to talk about it than average)

3 Predictive Task

As a point of reference to our models we designed a very simple predictive model that would function as baseline. Since "female" was the most frequent label our baseline would simply classify every row as "female", achieving a 39.01% accuracy.

We completed the task by using an ensemble of methods that would leverage on specific features in the dataset. We used three model in total:

- a convolutional neural network to analyze the profile pictures
- a gradient boosting classifier to classify the user behaviour and the description
- a multi-class classifier to classify the tweets of each user

In this section we will discuss in details the implementation of each model.

3.1 Inferring gender from user's behavior and description

For this model we had a total of 83 features that we used to train a Gradient Boosting Classifier (GBC). The GBC was chosen for its ability not to overfit (since we are averaging the results over multiple classifiers) and its readability (it is easy to tell which are the most important features and what is the decision process behind its classification). The first feature, the number of re-tweets of a user, was related to the users behavior and was provided directly by the original dataset without the need of pre-processing them. We extracted two features from the username of each user by returning the length of each username and by checking if males names (obtained from a list posted on the US Census website) were present in the username. We also had thirty topics distribution features obtained by applying LDA on the description of each user. Finally we had fifty features that represented the color of the sidebar of each user. Originally there were more then 700 colors. In order to reduce the dimensionality

of the color feature we encoded each color number into a one hot vector and applied PCA on the resulting matrix. It is important to mention also that the original dataset provided additional features, including the time zone of the user, the coordinates of the tweets etc. but we did not use them as there are not highly correlated to the gender of a user.

3.2 Inferring gender based on text from Tweets

3.2.1 Bag of Words

We implemented the bag of words model for gender classification. In the data for every twitter profile we had one of their tweet and also the description in their profile. We took the tweets from the profiles, removed punctuation symbols from them and converted all the words into lower case to ensure uniformity. Further, we calculated the frequency of processed words and selected the thousand most frequent words from the user tweets to fit the model. Below is a map for fifty most frequent words after processing. The top five words are - weather, im, get, channel and updates.

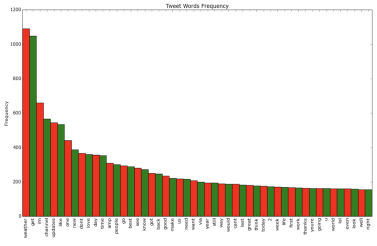

Fig 2: The top five words are - weather, im, get, channel and updates.

Similarly, we processed the words in the profile description and calculated the frequency of words to select the thousand most frequent words from profile description. The top five most frequent words are - love, im, news, life, music.

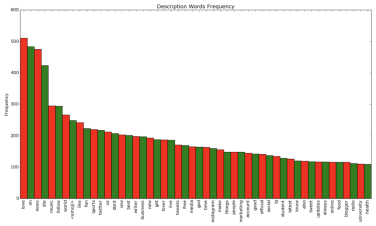

Fig 3: The top five most frequent words are - love, im, news, life, music.

Finally, after choosing top 1000 most frequent words from both tweets and description, we applied a multinomial naive bayes, one vs all classification to find probabilities of different profiles belonging to one of the classes - male, female or brand. The model performed fairly well as compared to the baseline estimation and we got an accuracy of 65 percent. We also analysed the most critical words in prediction for all the classes, based on the weights they got after training. Below are the maps for twenty most important words for each class.

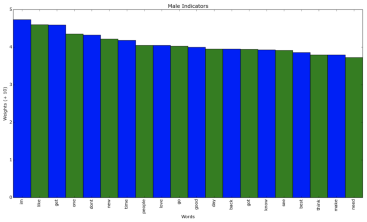

Fig 4: Maps for twenty most important words for each class.

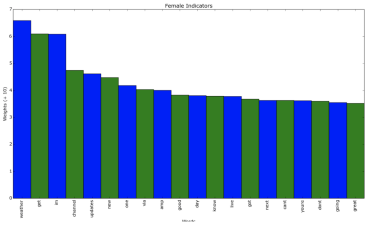

Fig 5: MFor female class, words like weather, im, get, channel, good etc were most positively weighted words.

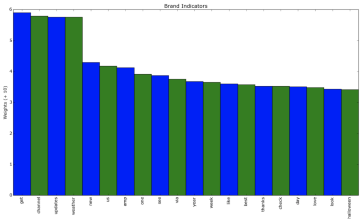

Fig 5: Similarly, for brands the most critical words in this model came out to be best, channel, new, week, see etc.Clearly brands dont use first person words much and the model correctly doesnt weigh them much.

3.2.2 Tf-Idf

In this model we selected the 1000 most frequent words from text and description similar to what we did in the bag-of-words approach but applied tf-idf instead of just relying on the frequency of the words because certain words like - im, get are used quite often and such words are not able to differentiate much between male and female profiles. Even though they work pretty well to distinguish gender and brands. We also included the sidebar-color of the profiles along with normal words as certain colors were predominant in male and female profiles and we hoped they will work well for the prediction task. This model again performed very well compared to the baseline task (in which we just assigned every profile the most dominant class in the dataset) and we got an accuracy of about 62 percent. Even though we were expecting this model to perform better than the bag-of-words model but it fell slightly shy as far as the accuracy is concerned. Further analysis of the most critical words for different classes in under this model gave following results.

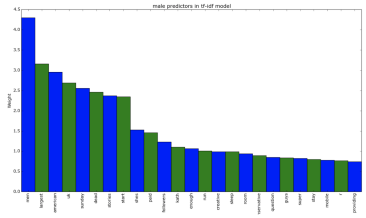

Fig 6: Top 5 Words for male - men, largest, american, uk, sunday

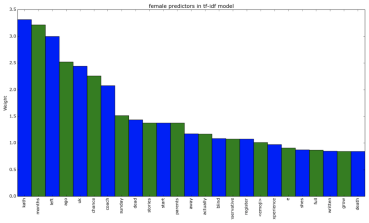

Fig 7: For females following words were important - months, left, age, emoji, stories etc.

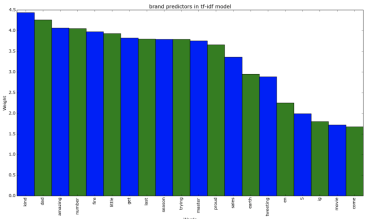

Fig 8: For brand prediction following words were most critical - amazing, number, get, last, season etc.

3.3 Deep Convolutional Neural Network for Image Classification

We are using VGG-19 architecture with transfer learning for upper fully connected layers. VGG-19 architecture is a 19 layered convolutional architecture. Convolutional networks (ConvNets) are currently considered state-of-the-art in visual recognition. We started with AlexNet which gave nearly 65% accuracy on test set, followed by VGG-16 which gave 67% accuracy on test set. We later settled to VGG-19 net as it was giving better accuracy (70%) on validation dataset and test dataset. We fixed lower layers and trained upper fully connected layers with three classes Male, Female and Brand for 20 epochs on training dataset. Our Initial observations regarding dataset were that some images were deleted, some were brand logos, some had pictures of couples, and rest were correct. We downloaded images using URLs and after some preprocessing like normalization and resizing each image to 50*50*3 we trained our image dataset on 80% of dataset and took 10% of dataset for validation and rest for testing. Some images even had 4th channel associated with them like alpha channel information, we completely ignored it because VGG nets only take 3 channel inputs.

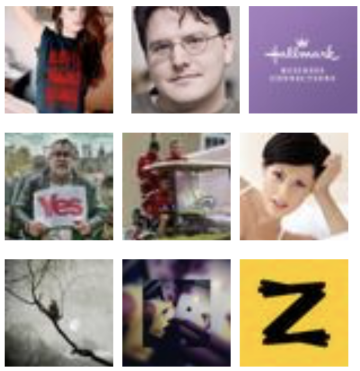

Fig 2: Samples of profile pictures from the dataset.

Major problems faced by us were that many people dont put their own picture on their profile. Some have put brand Images and some have put their childrens images. Some people have put images having both male and female like couple pictures. So our network could not learn the features differentiating the genders properly while training on these type of images. This is one of the reasons we were able to achieve only around 70% accuracy on our dataset based on input images only. But when these predictions were combined with text reasoning they gave a higher accuracy.

Fig 3: Architecture of VGG19

4 Model

We started by training the three models defined above using the same training dataset. Each model produced a softmax probability for each possible class where a higher value would correspond to a higher confidence towards a specific class.

Once we obtained the confidence of each model we combined the results in a new dataset consisting of a label and nine features. Each feature would correspond to the confidence of a specific model that a label was correct. For example $feature(model_i, brand) = P(brand = True|model_i)$.

We tried two different types of ensemble models. The first one, the simplest one, consisted in a weighted average. Each class probability was calculated by multiplying each model prediction with a weight proportional to their accuracy and by summing the results together. The class with the highest probability was then selected.

The second model was complex and it involved training a Gradient Boosting Classifier with the class predictions from each model in the training dataset and calculating the accuracy using the remaining validation set.

5 Results

Below we can observe the respective accuracies obtained using different models.

Model	Validation accuracy (%)
Baseline	39.01
Re-trained VGG19	67.00
GBC with LDA and num features	59.71
[]	
Ensemble model	86.47

Table 4: Accuracy from different models

	True "male"	True "female"	True "brand"
Pred "male"	901	24	27
Pred "female"	17	765	111
Pred "brand"	32	160	707

Table 5: Confusion matrix from the ensemble model)

As we can see from table 4 the ensemble method helps drastically to improve the accuracy. As previously shown by other projects as well, often combining models that leverage different features and strengths leads to more accurate models.

The final accuracy that we obtained is also far better than the initial baseline. We would also like to highlight that, as mentioned by Nguyen et al. [6], 10% of users don't employ language associated to their biological gender, making it harder to achieve an over 90% accuracy using text.

References

[1] Alowibdi, Jalal S., Ugo A. Buy, and S. Yu Philip. "Say it with colors: Language-independent gender classification on twitter." *Springer International Publishing*, 2014.

[2] Z. Miller, B. Dickinson and W. Hu, "Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features." *International Journal of Intelligence Science, Vol. 2 No. 4A*, 2012.

[3] Chen, X., Wang, Y., Agichtein, E., and Wang, F. "A Comparative Study of Demographic Attribute Inference in Twitter." *ICWSM, 15, 590-593*, 2015.

[4] Culotta, A., Kumar, N. R., and Cutler, J., "Predicting the Demographics of Twitter Users from Website Traffic Data." *In AAAI (pp. 72-78)*, 2015.

[5] Liu, W., and Ruths, D, "What's in a Name? Using First Names as Features for Gender Inference in Twitter.", *In AAAI spring symposium: Analyzing microtext (Vol. 13, p. 01)*, 2013

[6] Nguyen, D.; Trieschnigg, D.; Dogruo z, A. S.; Gravel, R.; Theune, M.; Meder, T.; and de Jong, "Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment." *In Proceedings of COLING* 2014