Raymond Hu

CIS787

**Final Report**

This project focused on twitter user gender classification to predict the gender of a person given their profile statistics and sample tweet. This is important because it allows people to get a sense of the user’s on twitter and analyze gender preferences. For example, if Twitter was classifying their own data, they would be able to use the results to better optimize their platform for each gender. From the data, we could predict what is trending among each gender, how active a person is given male or female, what the popular/most shared thing is, and the list can go on. Essential we can gather a lot of information about a person and in this case a specific gender based on their tweets, statistics like retweets, and profile customizations like color. This also an important business aspect for companies who want to target a specific gender so that for example an advertisement for tampons should be sent to classified females. Thus classifying gender on Twitter gives information on what each specific gender does on twitter.

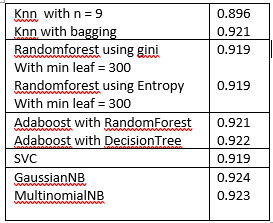
There have been plenty of twitter gender classifications papers and projects, but this is only going to mention relevant papers that fit. One prior work is a paper by Jalal Alowibdi that utilized empirical evaluations of profile without relying on text data like tweets for gender classification1. This paper achieved a 74% accuracy through NB classifier by just using color features of sidebar,link, and background color. To further improve color analysis, they mentioned “color quantization and sorting” which reduces the number distinct colors and sort it to get an understanding of what colors each gender typically uses. This is relevant to the project since it reduces color into binary attributes which makes it useable since either an instance has a color or not. One interesting thing was using name attribute to classify by cleaning it and breaking it down into n-grams up to 5 of characters and another one of phonemes. Processing name attributes tell us a lot about a user since a female user would probably use more feminine sounds. Another approach for names was utilizing a verified dictionary of female and male names to process twitter screen and user names by Marco Vicente2. Since twitter names can be modified and contain random extra characters, its solution to this was by removing characters and utilizing leet speak to modify the names until it matched a name in the dictionary. These two approaches to names would help break down a twitter name into parts that can be classified for gender prediction with good accuracy. For processing twitter text, Zachary Miller utilized n-grams up to 5 on the character level to present a tweet which created an immensely large feature space. One of the classifiers used was naïve bayes in which it just took the highest probability to determine if the instance was male or female given probability of the features created from the n-grams3. N-gram approach would help, but it would be more useful on the word-level for the project since char-level would take exponential amount of time & space that isn’t practical.

For this pre-processing stage, I first applied feature selection to filter out attributes that weren’t relevant to the task. Attributes like “unit\_id” or “profile\_yn” are either leftover by creator of dataset or hold no semantic value. For attributes missing a lot of values such as “tweet\_coord” which had over 90% null values got removed. Tweet location, timezone, and image URL were also removed because these attributes would have to utilize 3rd parties’ APIs to the links and places into useable data that would provide a good insight on the data. Furthermore, a simple glance at the data showed a lot of locations were not valid such as locations like “Worldwide” with around 1/3 of the records having null values. Image URL got removed since checking if each link contained a valid image and then processing that image conflicted with time constraints. For the attribute gender, 100 null instances got removed since I need the label for training purposes. Gender attribute also contained other distinct values besides male and female such as brand, unknown which got dropped reducing our total dataset size from 20050 to 12894 records. Lastly, the retweet\_count attribute got dropped since 2.6% of users within the 12894 had a count over 0 which isn’t useful and it essential is the same thing as having 97.4% missing data within a column.

For attributes regarding color, I used a package to convert the hexadecimal to rgb percentages, so it is constrained between 0 -1 to prevent any bias towards these attributes when classifying. I left any invalid hexadecimals as rgb value of 0,0,0 since replacing with mean has an actual meaning being a different color. Each color attribute is transformed into 3 new attributes of rgb values. For text pre-processing, each text record got symbols, numbers, and stop words removed and then lowercase and tokenized to get each word of the text. Then using each word, I applied 2 methods to clean it via lemmatization and stemming to get a lemmatized and stemmed version of the current word and added to a list depending if it was male or female. From these two lists I created a third list which was the intersection of both male and female sets which I subtract from both sets. This creates 2 pure male words and pure female words with the intersection set being words both genders use. The idea behind this way was that a person had a high count of pure male words with low female word count, then the person was probably a male. Then save all these sets of words in json. Lastly for text, generate the count of female, male , and intersect words for each text record for each method (lemmatization, stemming).

For name processing, each twitter name got symbols and numbers removed. Then generate 2 new attributes based on the number of vowels and constants in a name. The idea being, female names typically use more vowels such as ‘a’ for feminine sound. Based on the mean, female names typically use 0.5 more vowels. Final steps of pre-processing applied the log function to fav\_number and tweet\_count to reduce the range of the numbers since the maximum fav\_number was 341621 and tweet\_count was 2680199 which would cause a big bias when classifying relying on methods that use distance measure such as Euclidean.

For determining successfulness of the project, I plan to look at the accuracy of predictions from the classifier. For classification, I used 4 main types of classifiers being Knn, RandomForest, Bayes, and Svc. I utilize Knn clustering to determine optimally number of neighbors up to 10 and optimal test size in increments of 10% from 10% to 90%. The most optimal test size was 90% but to avoid underfitting , I used 60% test size and 40% training size. Using these sizes, I did holdout with other methods to compare accuracy rates. The table below accuracy rates for each method using 40% training data and 60% test data utilizing all attributes with 10 cross fold validation.

Ensemble methods such as bagging and boosting was utilize increase accuracy as shown in the table. Applying ensemble method added around 2% for Knn but with Adaboosting on Randomforest, the accuracy increase was practically unnoticeable at 0.2%. Surprisingly, Naïve Bayes performed the best even with independence assumption not holding true. Other experiments included using subsets of features for classification. Without utilizing text data from tweets, each classifier drops around 30% to around the 55%-62% range. Using stemming only or lemmatization of text only with all other attributes only provided a loss of 1% for each classifier. If we drop all other attributes besides text data, the model performs around the same as using all attributes. Other clustering methods were not used even though number of clusters is known to be 2 because to perform classification we need labels of points and clustering methods such as k-means doesn’t give you the class for each point.

**Citations**

1. Alowibdi, Jalal S., et al. “Empirical Evaluation of Profile Characteristics for Gender Classification on Twitter.” *2013 12th International Conference on Machine Learning and Applications*, 2013
2. Vicente, Marco, et al. “Twitter Gender Classification Using User Unstructured Information.” *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2015
3. Miller, Zachary, et al. “Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features.” *International Journal of Intelligence Science*, vol. 02, no. 04, 2012, pp. 143–148