**Computational Human Behavior Modeling**

**Assignment Due: February 27th, 2019; 11:59 pm**

In this homework, you will work with machine learning. You will build a classifier that can distinguish gender of a profile on social media. Your submission should include a (maximum) three-page **report** that discusses how you built the classifier, and then presents the performance of your classifier and also your **code**. The classifier should be based on features extracted from the training dataset. The format and content of the report is explained below.

**Dataset:**

Gender classifier data (made available in Assignment 2 zip file)

This data set was used to train a gender predictor. Contributors were asked to simply view a Twitter profile and judge whether the user was a male, a female, or a brand (non-individual). The dataset contains 20,000 rows, each with a user name, a random tweet, account profile and image, location, and even link and sidebar color.

**Classifier:**

You do not have to write a classifier from scratch. You are free to use one or more of the many open source (or other) tools and packages that allow you to use a variety of different classifiers. You can pick any package or programming language you like or are most comfortable with – although Python is the default language of most machine learning applications and a good thing to have on your resume.

Here is a list of available resources: (a) <https://textblob.readthedocs.org/en/dev/> (b) <http://www.cs.waikato.ac.nz/ml/weka/> (c) http://nlp.stanford.edu/software/

**Contents of the report (90 points + 10 points for working code and implementation of the algorithm):**

*1. Feature construction* (10 points). Like we covered in our lectures, constructing a classifier involves extracting relevant and meaningful features from the data under consideration. In your report, you will need to first present the various features you derived from the textual content of the data. Features can include (but not limited to) unigrams, bigrams, TF-IDF, Part-of-Speech tags, length of words etc. of the messages. Also, you will need to discuss why you chose the particular set of features. You can also use the image features, but you will need to download the images associated with the profiles to do so.

*2. Description of the classifier* (10 points). Discuss what is the particular classifier you chose (e.g., k Nearest Neighbor, Support Vector Machine, Naïve Bayes or some other), and a justification or rationale behind its choice and applicability to the dataset in question. That is, if you picked classifier X, why is it a good fit for this problem? Why is it a better choice compared to another classifier Y?

*3. Evaluation technique* (10 points). Present how you evaluated how well your classifier of choice performed in distinguishing gender. Particularly, discuss applicability of the concept of k-fold cross validation here. You will also need to present what metrics you used to evaluate performance of the classifier. For instance, typical metrics would include percentage accuracy, precision, recall.

*4. Implementation* (30 points).

a. Discuss how you preprocessed your data. If you used stopword removal, stemming, or tokenization over the content of the messages, you need to report it here. Point to the particular libraries or functions you needed for this.

b. Discuss how you extracted the features you presented above from the dataset. This needs to include what libraries and which particular functions you used for extracting each feature. If you did not use an existing library, you need to write about the method you used to compute the features from the data. Report if you did some filtering or feature selection to disregard not-so common features (e.g., if you ignored all unigrams which occurred less than five times). Also report if you did any kind of normalization or standardization of each feature, and your justification behind doing or not doing so.

c. Next, discuss how you implemented/used a library for your chosen classifier. Report what were the inputs and outputs to the particular library function you used and if/how you tuned parameters of the classifier (e.g., if you chose SVM, report the particular kernel you used).

d. Discuss how you partitioned the dataset for k-fold cross validation, along with what was your chosen k here. Here you will also discuss based on your chosen k-fold cross validation setup, what were your training and test sets in each of the k-iterations.

e. Discuss how you calculated the metrics of performance evaluation, e.g., accuracy, precision, recall etc. It is again okay to use an existing library that gives precision and recall values, in which case you need to present in your report which libraries/functions you used for the purpose, and what was your input and output to those functions.

*5. Analysis of results* (10 points). Report the performance of your classifier based on the above discussion. You will need to use charts, graphs, or tables to report actual numbers—i.e., the values of the 3 performance metrics you chose above (accuracy, precision, recall etc.). These numbers should be reported for each of the k iterations of the k-fold cross validation setup. You should also report the average performance over all k cross validation folds, corresponding to each evaluation metric.

*6. Applying classifier to conversational data* (20 points).

You have been given a dataset in Assignment 1. You will apply the classifier to the dataset and present the results and analyze the outcomes.