CIS 520 Final Report
Gender Classification using Twitter data
Team: Lagrange's Triads

Team Members:

1. Chirag Shah 2. Harshal Lehri 3.Prerna Srivastava

Models Tested

We tried out several algorithms using different features. The methods that worked out best are listed in the table below. In the end we went with an ensemble of SVM(only words), SVM(both words and image features) and RobustBoost

Index	Classifier Used	Features Used	Accuracy(10-Fold Cross Validation)
1	SVM (Kernel type : kernel_intersection)	Top 600 Words- normalized, and ordered in descending fashion as per the mean difference between male and female usage of the word	88-89%
2	SVM (Kernel type: kernel_intersection)	Top 600 Words- normalized, and ordered in descending fashion as per the mean difference between male and female usage of the word Image Features : age and smile also mean centered	88-89%
3	LDA (Linear Discriminant Analysis)	Top 600 Words- normalized, and ordered in descending fashion as per the mean difference between male and female usage of the word Image Features : age and smile also mean centered	88-89%
4	Logistic Regression	Top 200 normalized words, in descending order as per the mean difference between male and female usage of the word	85-86%
5	Linear Regression	PCA'ed words, 356	87-88%

		features selected that maintain 95% of the variance	
6	Adaboost (Robust Boost) and 1000 decision stumps	Top 600 words-normalized, mean-centered and ordered in descending fashion as per the mean difference between male and female usage of the word All image Features used	88-89%

Feature selection process:

First we began by analyzing the words given to us. After taking the mean across each class(male and female), we observed that many words were not at all used by both males and females while many words were used almost equally by males and females. These words would not really help in classifying between the 2 categories since they do not show any prominent or even minor difference between the two categories. This helped us reduce word features and increase accuracy by 3-4 percent.
We performed normalization on the words to equalize the word features based on usage. We then took the difference between the meal male and female word usage. By taking the absolute value of these features, we obtained the difference in usage of all the words between the two categories. The most prominent ones were obtained by sorting the mean differences in an descending manner. From the distributions of male and female image features, we found the first two to be most important in distinguishing the two and hence used them.
From the age distribution shown in can see in the above graph, the females belong to the younger age group as compared to the the males .Females have bigger smiles than males. Thus, the above features prove to be helpful in discriminating between male and females.

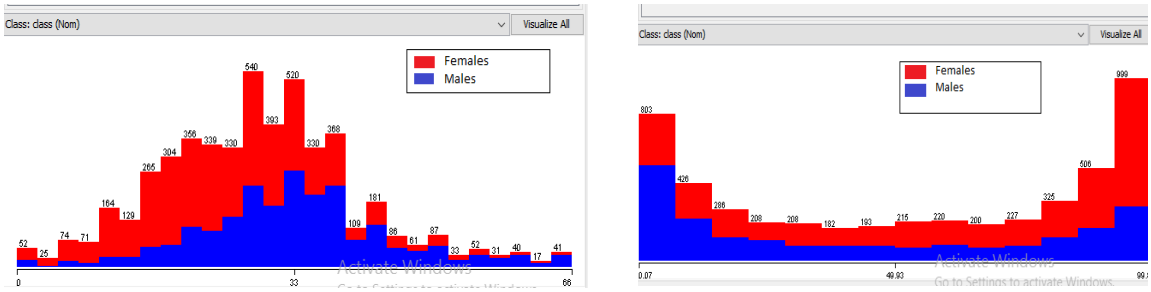

Figure 1 : Clockwise from top left - (a) Distribution of age in males and Females (b) Distribution of length of smile feature between males and female

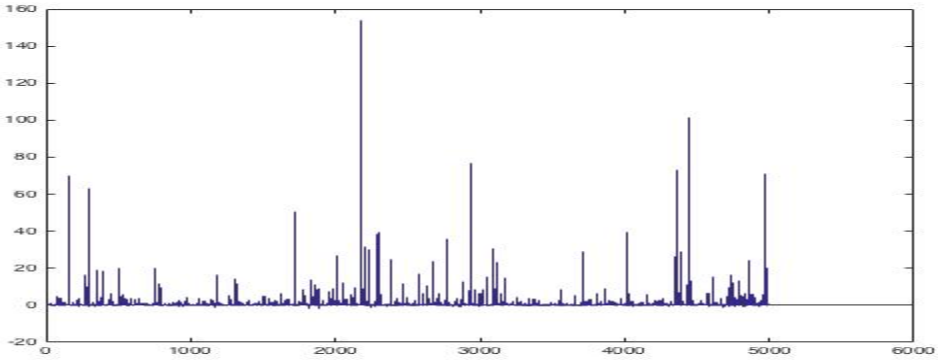

Figure 2 : Difference between (mean)male and (mean)female word usage. Some word are used more frequently by females than males however the reverse is not true.

The following table shows the frequency of times each algorithm correctly or incorrectly predicts the output.

We plot the difference between the two word distributions as shown. We see that there are some words which are used varyingly by male and female accounts and hence we consider only the top 600 words whose mean “usage” for male and female samples is large.Our final ensemble consisted only of three algorithms, RobustBoost, SVM on words and another SVM on image features and words.

SVM(only words)	**SVM(image features and words)**	**AdaBoost(image features and words)**	**Frequency(Average over 10 iterations)**
Incorrect	Incorrect	Incorrect	53.3
Incorrect	Incorrect	Correct	30
Incorrect	Correct	Incorrect	9.5
Incorrect	Correct	Correct	30
Correct	Incorrect	Incorrect	14.7
Correct	Incorrect	Correct	20.1
Correct	Correct	Incorrect	56.5
Correct	Correct	Correct	783.9

It can be seen that on average all three algorithms correctly predict the result for classification or two of the others correct the mistakes of one of them. However there are a few cases where the ensemble predicts incorrectly though one of them is correct. We also see some examples which confuse all three classifiers.

To see the mean values of features that confuse the classifiers, we plot a graph showing the mean of the values of the samples that the algorithms classified incorrectly.

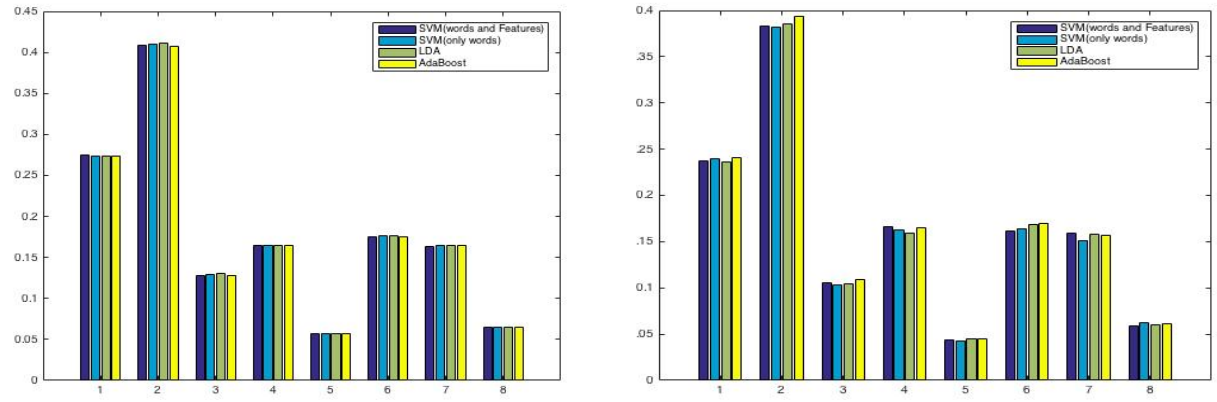

Figure 3 : (a) The mean value of first 8 word features(ordered and normalized as described above) of samples that are correctly predicted by the algorithms. (b) The mean value of the first 8 word features(ordered and normalized as described above) of samples incorrectly predicted by the algorithms.

From the above graph, it can be seen that there is very less difference between the mean values that of the samples incorrectly classified by the algorithms and the mean values of the samples correctly classified by them. It can also be observed that the mean value of a feature misclassified(or correctly classified) is about the same for all four methods. Hence this ensemble increased the overall accuracy only by a small percent(to 90%). Due to space limitations we were unable to take Linear Regression into account. We also noticed that SVM's were more biased to classify an example as female but that might be due to there being more women in the training set.

Derived Features:

We tried extracting some new features from words and images. The derived word features were: average word length, total number of words used, number of unique words and vocabulary richness (number of unique words/total number of words) for each user. Running classification only on these features gave an accuracy of 70%. Using these along with the primary features (words and image features) did not provide much increase in the accuracy of the base models. From the graphs we can see that these features had the same distribution for males and females, specially, average number of words used, for both the males and females, where the graph has the same distribution peaking at around 3.8. That is why including these did not help much.

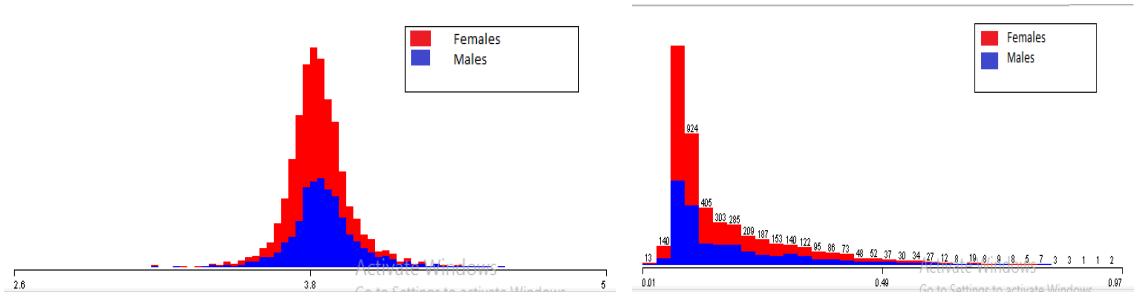

Figure 4: (a) Distribution of average word length for males and females (b) Distribution for vocabulary richness for males and females

The various features that we derived from images were HOG and SIFT features. Using these alone as features to the classifiers again gave an accuracy of around 70%. This the face images were not distinct and there was a lot of background noise in the images. We tried using PCA on the images to detect eigenfaces as well but it did not perform too well because of the background. These could have worked with a combination with other features (word and image features) to boost the accuracy of the other classifiers. Another feature on images we tried were the HELO features(Histogram of edge Local Orientations) that creates a histogram of gradients of all points in the image. However that too alone was giving an accuracy of 70%.

We also noticed that some ages in the training set were from 0-12 which were incorrect. Hence we tried to cluster similar age groups based on the word distributions of each age group. However this did not increase the accuracy by much. Age binning also did not improve the accuracy by much. The clustering(number of clusters =6) of age groups is shown below:

Cluster ID	1	2	3	4	5	6
Age Groups	0,10,12,13, 14,15,16,17, 18,20,21,22, 23,24,25,26, 27,28,29,30, 31,32	61	1,56,62	2,4,5,6,7,8, 9,11,19	2,33,34,35, 36,37,38,39, 40,41,42,43, 44,47,49,51, 52,53,55	45,46,48,50, 54,57,58,59, 60,63,64,65, 66

Other things we tried:

We tried mixing the test data with the train data. For this, we took an ensemble of 7 methods and only selected the indexes where all 7 methods agreed. They agreed on 3637 decisions of the total 4997 decisions to be made on the test data. The accuracy of training the model using this data and testing on training data was 89.14%. According to these results, the model which was the combination of train data and selected test data should have generalized the data in a better manner. Though when we tested these models on the leaderboard, our overall accuracy decreased from 90.13 to around 89.1 . Hence this might have been due to overfitting and we did not submit this model.

We also tried using KLDivergence to find similarity of word distributions. However since both distributions had some words not present in the other, some values of KLDivergence were negative and hence did not give good results. Other distance metrics we used were the chi-square distance and Bhattacharya distance. But intersection kernels gave best results.

Key takeaway:

1. Think Simple. Do not try and overthink the problem. Parameter tuning is more important than using different models not tuned properly.
2. Theoretical foundations and empirical observations may not always match. The main reason for this is that theoretical foundations are usually based on assumptions which might not conform with the real life data.
3. Be wary of overfitting. More data does not always mean better generalization. The data has to be reliable.