CIS 520 Project Final Report

Woodpecker (Xiang Deng, Yiren Lu, Dongni Wang)

Fall 2015

For the final project, we developed a system for gender prediction (male/female) from the language of their tweets and the image they post with their twitter profile. We were given a training set of 4998 labeled training samples and a testing set of 4997 testing samples. Each sample has 5000 words features, 7 pre-extracted image features and 30000 raw RGB image pixel features.

In our system, we used seven classifiers on different feature sets and combined them using the stacking method. The seven classifier are: a logistic regression model on words features, an ensemble model consists of 300 decision stump trees using LogitBoost on selected words and image features, a SVM model with intersection kernel on selected words and image features, a SVM model with intersection kernel on selected and normalized words and image features, an ANN model with 2 hidden layers each with 100 and 50 nodes, a SVM model with RBF kernel on PCA-ed HOG features on face-detected images, and a SVM model with RBF kernel on PCA-ed LBP features on face-detected images. For the stacking method, we took the raw outputs (probabilities) from the seven basic models mentioned above trained with 80% of training samples and trained a logistic regression model using the other 20% of training sample. Our final full model achieved an overall accuracy of 92.42%. In order to meet the time and space constraint for the competition, we dropped the SVM model on PCA-ed LBP features and replaced the SVM model on PCA-ed HOG features with one bagging of logistic regression classifiers on raw HOG features. The submitted model for final competition achieved an accuracy of 91.04%.

In the following sections, we present the cross-validation accuracies of each methods we tried and discuss the rationale of our final model. We also provide some interesting visualization such as the most predictive words and the visualization of auto-encoder.

1 Methods

In this section, we report the results of multiple methods we tried for feature extraction, dimension reduction, and classification.

1.1 Data preprocessing

1.2 Feature Selection

To extract features from the raw word and image features, we experimented with multiple feature selection methods, including Information Gain, BNS

1.3 Dimension Reduction

1.4 Classification

2 Experiment Analysis

In this section, we analyze the results of our experiments of multiple methods for feature extraction, dimension reduction, and classification.

2.1 Feature Selection/Extraction

2.2 Dimension Reduction

2.3 Classification

The table of approaches and their associated 5-fold cross-validation classification accuracies are shown in the Table 1

Feature	Approach: Dimension Reduction	Approach: Classifier	Accuracy (%)
Words + Image features	PCA(500)	Ridge Regression + Sigmoid	$\approx 70\%$
Words + Image features	PCA(320)	Ridge Regression + Sigmoid	$\approx 79\%$
Words + Image features	PCA(2000)	Logistic Regression	$\approx 85\%$
Words	None	Logistic Regression	85.96%
Words	IG(1000)*	Logistic Regression	85.79%
Normalized-Words	None	Naive Bayes	72.25%
Words	IG(100)	multinomial Naive Bayes	77.95%
Words	None	Bernoulli Naive Bayes	79.59%
Words	IG(350)	Bernoulli Naive Bayes + EM	82.49%
Words	PCA(2000)	Artificial Neural Network	$\approx 86\%$
Words	IG(76)	K-Nearest Neighbor (L2)	72.89%
Words	IG(84)	K-Nearest Neighbor (Minkowski)	71.43%
Words	IG(95)	Random Forest	83.32%
Words	None	K-means	$\approx 60\%$
Words	IG(1000)	Decision stumps + LogitBoost	89.11%
Face-detected Image RGB	PCA(100)	Random Forest	$\approx 69\%$
Face-detected Image RGB	Auto-encoder(100)	Logistic Regression	75.17%
Raw HOG features over Face-detected Image RGB	None	Logistic Regression	$\approx 80\%$
Raw HOG features (Face/eyes/ nose-detected Image RGB)	None	Logistic Regression	$\approx 81\%$
Raw HOG features (Face/eyes /nose-detected Image RGB) + Gaussian Pyramid	None	Logistic Regression	$\approx 82\%$
Row HOG features (Face/eyes /nose-detected Image RGB) + Gaussian Pyramid	None	SVM (RBF kernel)	$\approx 83\%$
HOG features (Face/eyes /nose-detected Image RGB) + Gaussian Pyramid PCA(1500)	None	SVM (RBF kernel)	$\approx 84\%$
Dense LBP (Face-detected Image RGB)	None	SVM (RBF kernel)	$\approx 85\%$
*IG represents Information Gain			

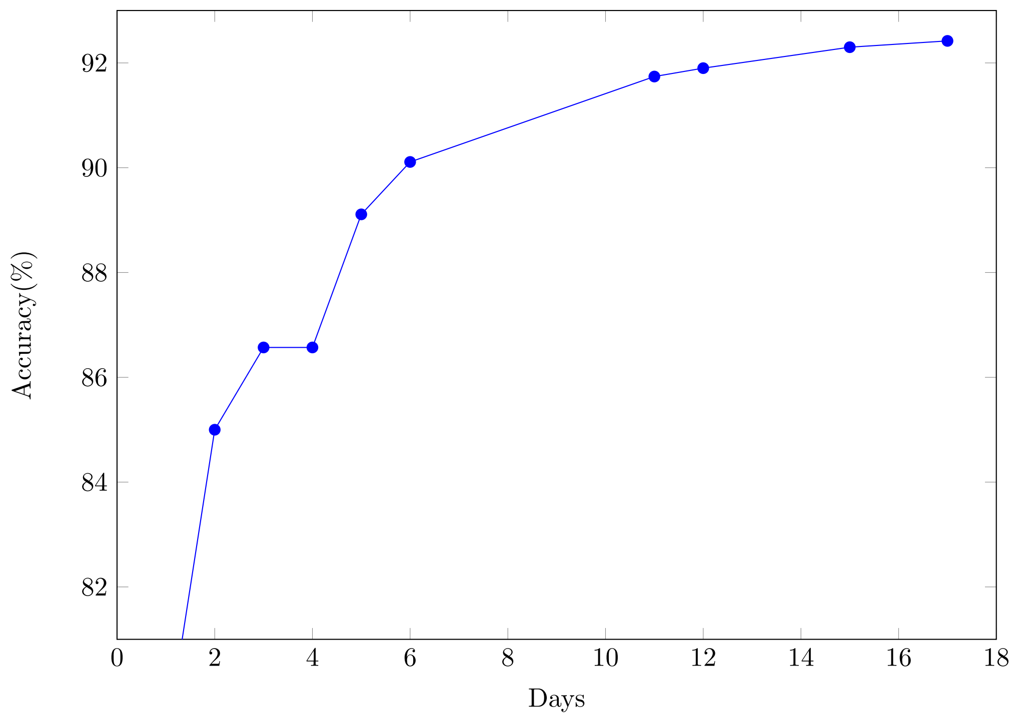
Table 1: Experimental results of single classifiers

3 Visualization

Approach		
Preprocessing	Classifier	Accuracy (%)
Features: Words + Image features		
IG(1000) on words and image features	Logistic (W) + Neural Network (W) + Ensemble trees (W+I) + stacking	91.11%
IG(1000) on words and image features	Logistic (W) + Neural Network (W) + Ensemble trees (W+I) + cascading	89.04%
Features: Words + Image features + Image RGB		
IG(1000) on words and image features; Face-detected image RGB PCA(100)	Logistic (W) + Neural Network (W) + Ensemble trees (W+I) + Logistic (PCA-ed RGB) + stacking	90.37%
IG(1000) on words and image features; Face-detected image HOG features	Logistic (W) + Neural Network (W) + Ensemble trees (W+I) Logistic (HOG) + stacking	91.55%
IG(1000) on words and image features; Face/eyes/nose-detected image HOG features	Logistic (W) + Neural Network (W) + Ensemble trees (W+I)+ Logistic (HOG) + stacking	91.74%
IG(1000) on words and image features; HOG features (Face/eyes /nose-detected Image RGB) Gaussian Pyramid	Logistic (W) + Neural Network (W) + Ensemble trees (W+I) Logistic (HOG) + stacking	91.9%
IG(1000) on words and image features; HOG features (Face/eyes /nose-detected Image RGB) Gaussian Pyramid + PCA(1500) Normalization	Logistic (W) + Neural Network (W) + Ensemble trees (W+I) Logistic (HOG) + stacking (SVM)	92.3%
IG(1000) on words and image features; HOG features (Face/eyes /nose-detected Image RGB) Gaussian Pyramid + PCA(1500) + Dense LBP + Normalization	Logistic (W) + Neural Network (W) + Ensemble trees (W+I) Logistic (HOG) + stacking (SVM)	92.42%
*For the classifiers: W: words; I: Image features		
*For the stacking method, we use logistic regression if not specified		

Table 2: Experimental results of ensemble classifiers

In this section, we include some interesting visualization obtained during the process of analyzing data, training, tuning, and testing our models.



4 Discussion

Working on this gender-classification project gave our team a chance reflected on what we have learned in class. Here is a short summary of things that have surprise us (or have taught us a lesson)

- With different feature sets (especially they have various ranges and dimensions), feature selection and normalization have played an important role in improving the performances of our model.
- Last but not least, be careful with required formats..