Discriminating Gender on Twitter

This paper approaches gender classification as a binary problem, ie you are either male or female.

Data

Data was gathered by using Twitters API, retrieving 400’000 tweets a day. Dataset was 213 million tweets, by 18.5 million users.
Twitter Profile has following characteristics: Screen Name(Mandatory), Full Name, Location, URL and Description
For ground truth, the blogs attached to 184000 twitters users were followed to get their gender
A subset of this number were manually checked to ensure selected gender was accurate
Simple analysis: Women tweet more, and high tweet rates correspond to having a blog
Automatic language translation was used for non English tweets

Features

Used Features: Screen Name, Full Name, Description, and Tweet
Descriptions and such could have changed over time.
Table 4 gives overview of features. This answer describes what N-grams are.

Experiments

As the database is so large, each feature pattern is translate to integer code and all users features are concatenated together into one vector.
Initial Results: Balanced Winnow2 (74%)>LibSVM(71.8%)>Naive Bayes(67%)
Learning Parameters was low for one feature, high for multiple features
Three main conditions used: Single Tweet, All Tweets, 4 Fields
Amazon Mechanical Turk was used for Human Performance

Results

Baseline (F)	54.9%
One tweet text	67.8
Description	71.2
All tweet texts	75.5
Screen name (e.g. *jsmith92*)	77.1
Full name (e.g. *John Smith*)	89.1
Tweet texts + screen name	81.4
Tweet texts + screen name + description	84.3
All four fields	92.0

Condition	Train	Dev	Test
Baseline (F)	54.8%	54.9	54.3
One tweet text	77.8	67.8	66.5
Tweet texts	77.9	75.5	74.5
All fields	98.6	92.0	91.8