Inferring Gender From the Context of Tweets: A Region Specific Example

Uses a subset of users from Nigeria, contains both a machine and non-machine learning approach
Determines gender from tweet content alone.

Data Collection and Preparation

Data was collected using Twitter's Search API with Geocode.
45 Nigerian cities with population > 100000 with 40 miles radius
Name gender pairs were obtained by searching name on Facebook, which has a gender field. This provided ground truth for twitter users gender
Any Non English names were translated
Left with 11155 users with 18.5 million tweets
Emoticons and symbols were translated to their meaning

Feature Selection and Machine Learning Experiments

Features were generated for unigrams, hashtags and psychometric properties.
Linquistic Inquiry and Word Count was performed on each tweet. This maps certain high frequency words to psychometric properties.
Unigrams features included all words but stop words
Hashtags contained the 75% most common ones
SVMLight was used with various combinations of these feature vectors
SVMLight is an SVM classifier with a linear kernel
80/20 training to test ratio

Feature Class	Prec.	Rec.	F	Accu.
Hashtag	70.82	46.99	56.48	63.81
Hashtag, LIWC	71.51	47.17	56.83	64.18
Hashtag, LIWC, Unigram	**82.53**	**77.64**	**80.00**	**80.60**
Hashtag, Unigram	82.06	77.46	79.69	80.26
LIWC	70.76	75.07	72.85	72.02
LIWC, Unigram	82.22	77.79	79.93	80.48
Unigram	**82.50**	**77.47**	**79.90**	**80.51**

Table 2 – Classification Results