Discriminating Gender on Twitter

Authors: John D. Burger, John Henderson, George Kim, and Guido Zarrella

Presented by: Komal Narwekar

Content

1. Introduction
2. Purpose
3. Data Sets
4. Features
5. Experiments
6. Results
7. Conclusion
8. Future Study

Introduction

- Demographic Research Interest in Social Media

Gender

Age

- Purpose:

Marketing

Personalization

Legal Investigation

Purpose

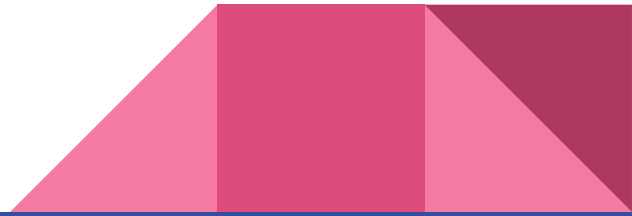
Identifying gender of twitter user.

Binary Classification Problem

Previous Work:

Manual Annotation on a dataset of 500 English users labeled with gender.

Feasibility?



Data- Twitter Statistics

In late 2010, it was estimated that Twitter had 175 million registered users worldwide, producing 65 million tweets per day.

Screen name (Mandatory)

Full name

Location

URL

Description

DataSet

Raw Data- 213 million tweets from 18.5 million users.

Preprocessing- Linking user profiles to their blogging website and extract gender.

- Filtered out spammers.

Final DataSet- 184,000 Twitter u

Quality Assurance Study

55% females and 45% males.

	Users	Tweets
Training	146,925	3,280,532
Development	18,380	403,830
Test	18,424	418,072

Figure 1: Dataset Sizes

Features

Features Used:

- Full Name
- Screen Name
- Tweet Text
- Description

	Feature extraction		
	Char ngrams	Word ngrams	Distinct features
Screen name	1–5	*none*	432,606
Full name	1–5	1	432,820
Description	1–5	1–2	1,299,556
Tweets	1–5	1–2	13,407,571
Total			15,572,522

Figure 4: Feature types and counts

Experiments

-Machine Learning Tools Considered-WEKA or MALLET

-Problems- Huge Data Size

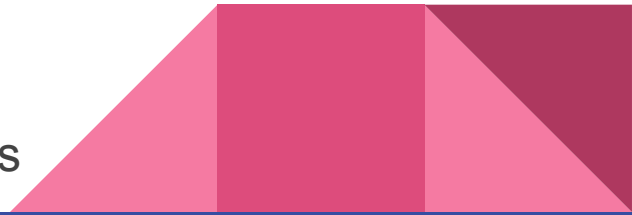
-One Time Preprocessing- Convert each feature pattern to an integer codeword.

-Classifier Used- Balanced Winnow2 with customization to deal with large data

Parameters Used:

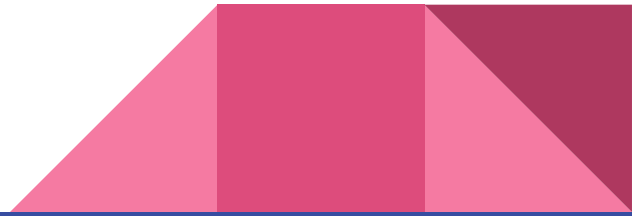
Low Learning Rate (0.03) - Single Feature

Higher Learning Rate (0.20) - Combination of Features



Experiments

- Field combinations-
  - Single Tweet
  - All Fields
  - All Tweets.
- Human performance
  - Amazon Mechanical Turk (AMT)
  - Simple Majority Vote
  - Expectation Maximization Algorithm



Results- Field Combinations

Prominent Figures:

- Best accuracy for all four fields: 92%
- User's full name: 89.1%
- Screen Name: 77.1%
- Tweets convey more about a Twitter user's gender than their own self-descriptions: (75.5% vs. 71.2%).
- Combination of Tweets, Screen Name, and Description: 84.3%

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

Figure 5: Development set accuracy using various fields

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

Figure 6: Accuracy on the training, development and test sets

Results- Field Combinations

Performance vs. Training Data Size

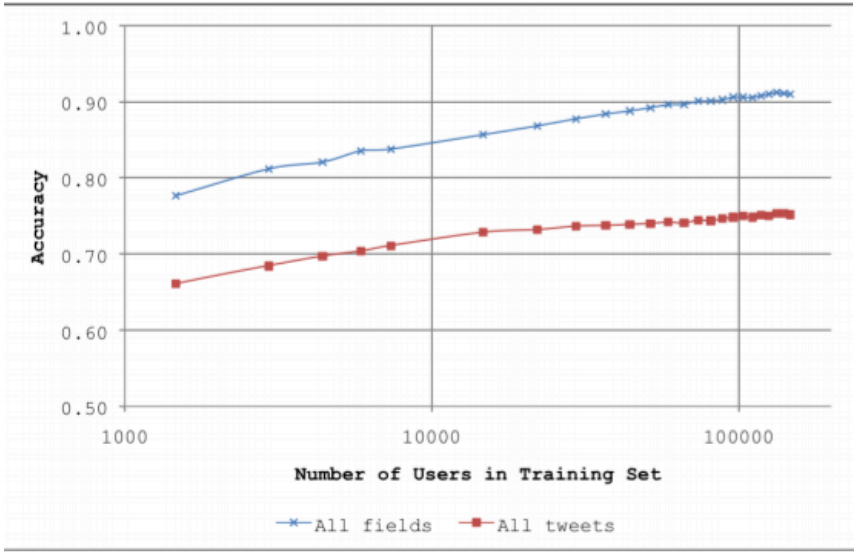
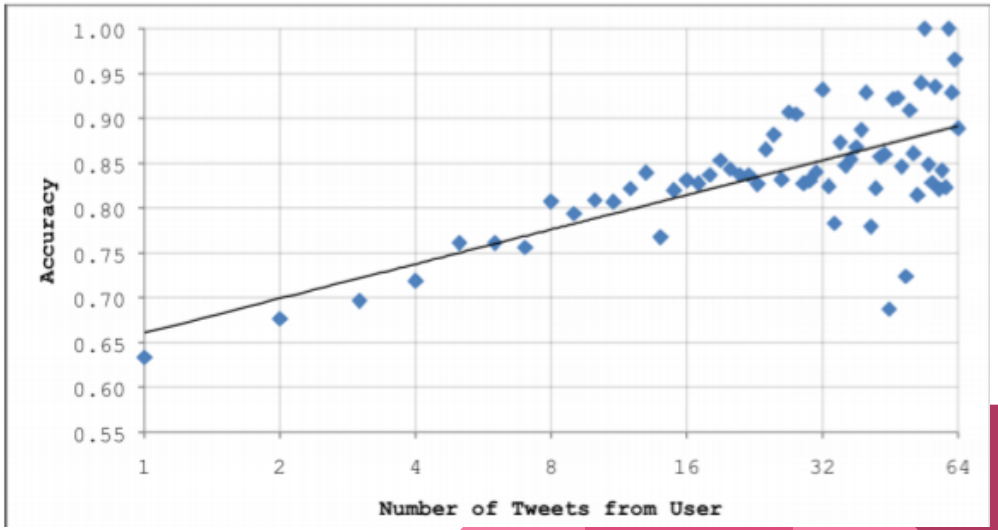


Figure 8: Performance increases when training with more users



Results- Field Combinations

Features that strongly convey gender:

Only 5 of the top 1000 features are associated more strongly with males.

Lower P(Female|feature) than the prior, P(Female) = 0.55.

Rank	MI	Feature f	$P(Female\|f)$
1	0.0170	!	0.601
2	0.0164	_:	0.656
3	0.0163	_lov	0.687
4	0.0162	love	0.680
5	0.0161	lov	0.676
6	0.0160	_love	0.689
7	0.0160	!_	0.618
8	0.0149	:)	0.697
9	0.0148	y!	0.687
10	0.0145	my	0.637
11	0.0143	love_	0.691
12	0.0143	haha	0.705
13	0.0141	my_	0.634
14	0.0140	_my	0.637
15	0.0140	_:)	0.697
16	0.0139	_my	0.634
17	0.0138	!_i	0.711
18	0.0138	hah	0.698
19	0.0137	_hah	0.714
20	0.0135	_so	0.661
21	0.0134	_haha	0.714
22	0.0132	so	0.661
23	0.0128	_i	0.618
24	0.0127	ooo	0.708
25	0.0126	!_i	0.743
26	0.0123	i_lov	0.728
27	0.0120	ove_	0.671
28	0.0117	ay!	0.718
29	0.0116	aha	0.678
30	0.0116	<3	0.856
31	0.0115	_cute	0.826
32	0.0114	i_lo	0.704
33	0.0114	:)$	0.701
34	0.0110	:(	0.731
35	0.0109	_:)$	0.701
36	0.0109	!$	0.614
37	0.0107	ahah	0.716
38	0.0106	_<3	0.857
464	0.0051	_ht	♂ 0.506
465	0.0051	hank	0.641
466	0.0051	too_	0.659
467	0.0051	_yay!	0.818
468	0.0051	_http	♂ 0.506
469	0.0051	_htt	♂ 0.506
624	0.0047	Googl	♂ 0.317
625	0.0047	ing!_	0.718
626	0.0047	hair_	0.749
627	0.0047	_b	0.573
628	0.0047	y_:	0.725
629	0.0046	Goog	♂ 0.318

Results- Human performance

Raw per-response performance is 60.4%, only modera
better than the all-female baseline.

Results with Majority Vote: 65.7%

Most workers perform below 80% accuracy

Less than 5% of the prolific workers out-perform the
automatic classifier.

Conclusion: Automatic classifier performs as well or better
than the AMT workers on their subset.

Baseline	54.9
Average response	60.4
Average worker	68.7
Average worker (100 or more responses)	62.2
Worker ensemble, majority vote	65.7
Worker ensemble, EM-adjusted vote	67.3
Winnow all-tweet-texts classifier	75.5

Figure 10: Comparing with humans on the all tweet texts task

Results- Human performance

System vs. Human

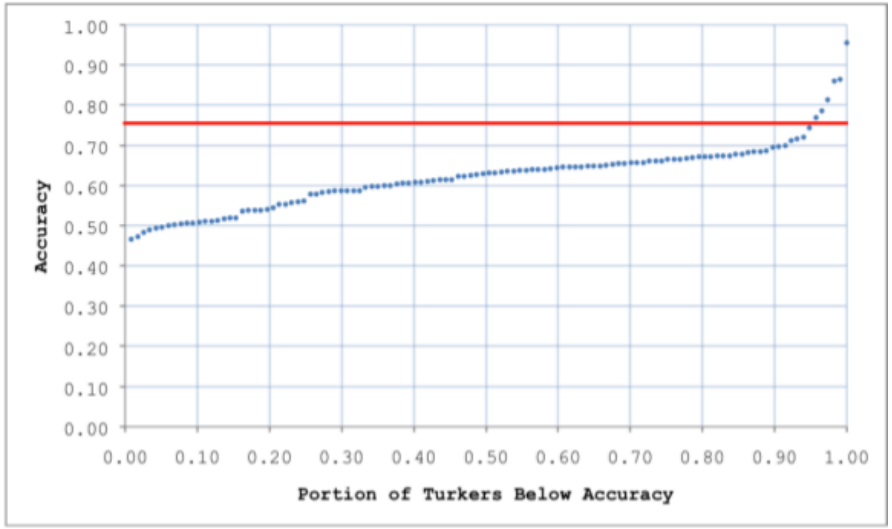
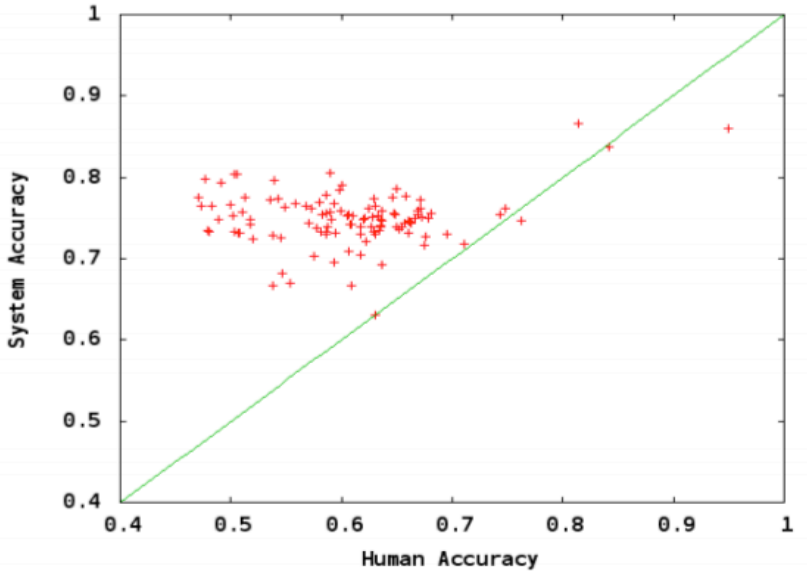


Figure 11: Human accuracy in rank order (100 responses or more), with classifier performance (line)

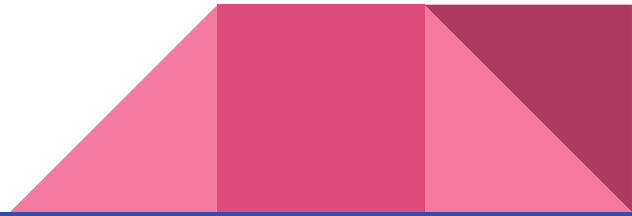


Results- Self-training

Results in a drop to 91.1% by training only Half Data Set

With the newly introduced Label Errors, performance drops to 90.9%

Solution Proposed: Using larger amounts of unsupervised data

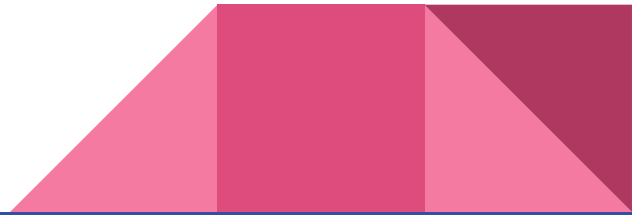


Conclusion

Best classifier performed at 92% accuracy- (All Fields)

Only tweet texts performed at 76% accuracy.

Only 5% of 130 humans performed 100 or more classifications with higher accuracy than this machine



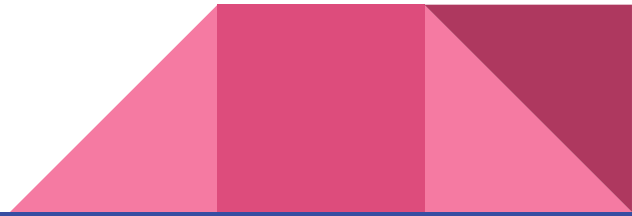
Future Study

Gender Identification in other Informal Online Genres:

- Chat Rooms
- Forum Comments

Assign other Demographic Features:

- Age
- Location



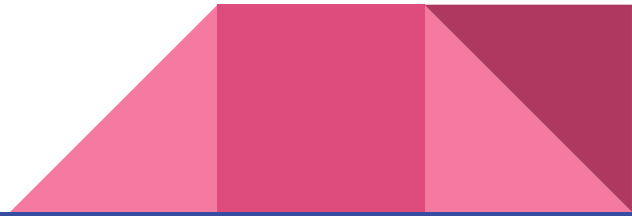
My Views

It is a fairly exhaustive classifier with support for various features.

Better accuracy than humans.

n-gram implementation can support various languages.

Could be much more useful if the self training model was developed further.



Questions

1. What other feature extraction methods the authors could have used to save themselves from the trouble of huge feature vectors?.
2. Does features with blank fields should be considered for training? How will they help in learning? If they do not help in learning, should they be considered for pruning?.
3. This method doesn't account for the fake profiles from the dataset on twitter as it may affect the accuracy of predicting the gender?
4. What is the need to compare classifiers efficiency with human performance?
5. Authors used training dataset to build classifier for unlabeled data so what could be the chances that this will give us same accuracy for data which is not similar to training set?
6. The subset of Twitter users who also use a blog site may be different from the Twitter population as a whole,so how would the sample size be representative in authors' approach?
7. Does using word -level ngrams give better results than using character -level ngrams?
8. What are other characteristics or parameters that you think authors could have included in their approach to make it more effective?
9. Will accuracy improve by combining sociolinguistic -based features with this approach?
10. Is crowd-sourcing a viable option for other Social Computing tasks?

Thank You