

WHAT'S IN A NAME?

Using first names as features for gender inference in Twitter

Wendy Liu and Derek Ruths
School of Computer Science, McGill University

First names carry signal.

First names carry signal.

How can we use this signal?

First names carry signal.

How can we use this signal?

What are the limitations?

Our investigation

First name — gender

Prior work: feature-based classifiers

Burger, J.; Henderson, J.; Kim, G.; and Zarrella, G. 2011. Discriminating Gender on Twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Pennacchiotti, M., and Popescu, A. 2011. A machine learning approach to Twitter user classification. In Proceedings of the International Conference on Weblogs and Social Media.

Name-based classifiers

vs.

baseline classifier

But first, we needed a dataset.

But first, we needed a dataset.

No canonical gender-labelled dataset

To avoid: deriving labels from text.

To avoid: deriving labels from text.

Common practice

To avoid: deriving labels from text.

Common practice

Inflates accuracy

Username

Latest status message

Gender?

kelseygreenwell

I can't get over how perfect my prom dress is! 😊

OfficeOfSteve

I would take pictures of myself at the gym, but I'm afraid I'd lose my man card on Twitter when people see that I bench press 40lbs

Zatics

I don't think I'll ever understand why people are obsessed with Nutella

Username

Latest status message

Gender?

kelseygreenwell

I can't get over how perfect my prom dress is! 😊

OfficeOfSteve

I would take pictures of myself at the gym, but I'm afraid I'd lose my man card on Twitter when people see that I bench press 40lbs

Zatics

I don't think I'll ever understand why people are obsessed with Nutella

Username

Latest status message

Gender?

kelseygreenwell

I can't get over how perfect my prom dress is! 😊

Female

OfficeOfSteve

I would take pictures of myself at the gym, but I'm afraid I'd lose my man card on Twitter when people see that I bench press 40lbs

Zatics

I don't think I'll ever understand why people are obsessed with Nutella

Username

Latest status message

Gender?

kelseygreenwell

I can't get over how perfect my prom dress is! 😊

Female

OfficeOfSteve

I would take pictures of myself at the gym, but I'm afraid I'd lose my man card on Twitter when people see that I bench press 40lbs

Zatics

I don't think I'll ever understand why people are obsessed with Nutella

Username

Latest status message

Gender?

kelseygreenwell

I can't get over how perfect my prom dress is! 😊

Female

OfficeOfSteve

I would take pictures of myself at the gym, but I'm afraid I'd lose my man card on Twitter when people see that I bench press 40lbs


Male

Zatics

I don't think I'll ever understand why people are obsessed with Nutella

Username	Latest status message	Gender?
kelseygreenwell	I can't get over how perfect my prom dress is! 😊	Female
OfficeOfSteve	I would take pictures of myself at the gym, but I'm afraid I'd lose my man card on Twitter when people see that I bench press 40lbs	Male
Zatics	I don't think I'll ever understand why people are obsessed with Nutella	

Username	Latest status message	Gender?
kelseygreenwell	I can't get over how perfect my prom dress is! 😊	Female
OfficeOfSteve	I would take pictures of myself at the gym, but I'm afraid I'd lose my man card on Twitter when people see that I bench press 40lbs	Male
Zatics	I don't think I'll ever understand why people are obsessed with Nutella	?

Username	Latest status message	Gender?
kelseygreenwell	I can't get over how perfect my prom dress is! 😊	Female
OfficeOfSteve	I would take pictures of myself at the gym, but I'm afraid I'd lose my man card on Twitter when people see that I bench press 40lbs	Male
Zatics	I don't think I'll ever understand why people are obsessed with... 	?

Result: cherry-picking users

Excludes users that are difficult to categorise

Our approach: profile pictures.



Amazon Mechanical Turk



Amazon Mechanical Turk

20 users per task



Amazon Mechanical Turk

20 users per task
3 workers per user



Result: 12,681 labelled users.

4,449 male, 8,232 female

Result: 12,681 labelled users.

4,449 male, 8,232 female

Download from bit.ly/microtext2013

The classifier

Support vector machine

Prior work:

Zamal, F. A.; Liu, W.; and Ruths, D. 2012. Homophily and latent attribute inference: inferring latent attributes of Twitter users from neighbors. In Proceedings of the International Conference on Weblogs and Social Media.

Liu, W.; Zamal, F. A.; and Ruths, D. 2012. Using social media to infer gender composition from commuter populations. In Proceedings of the When the City Meets the Citizen Workshop, the International Conference on Weblogs and Social Media.

Features

k-top

words ("hello")

digrams ("he", "el", "ll", "lo")

trigrams ("hel", "ell", "llo")

stems ("hel")

co-stems ("lo")

hashtags ("#hello")

} Lovins stemming algorithm

frequency (number per day)

tweets, mentions, hashtags, links, retweets

ratios

tweets to retweets

followers to followees

SVM kernel

radial basis function

parameters chosen using grid-search

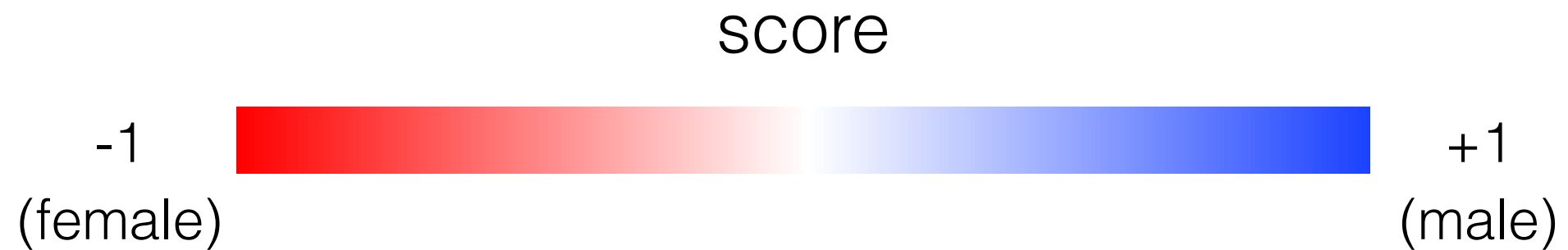
US census data (1990)

name

score



$$\frac{(\text{Number of males with this name}) - (\text{Number of females with this name})}{(\text{Number of people with this name})}$$



Three gender inference methods

Three gender inference methods

Baseline

Three gender inference methods

Baseline

Integrated

Three gender inference methods

Baseline

Integrated

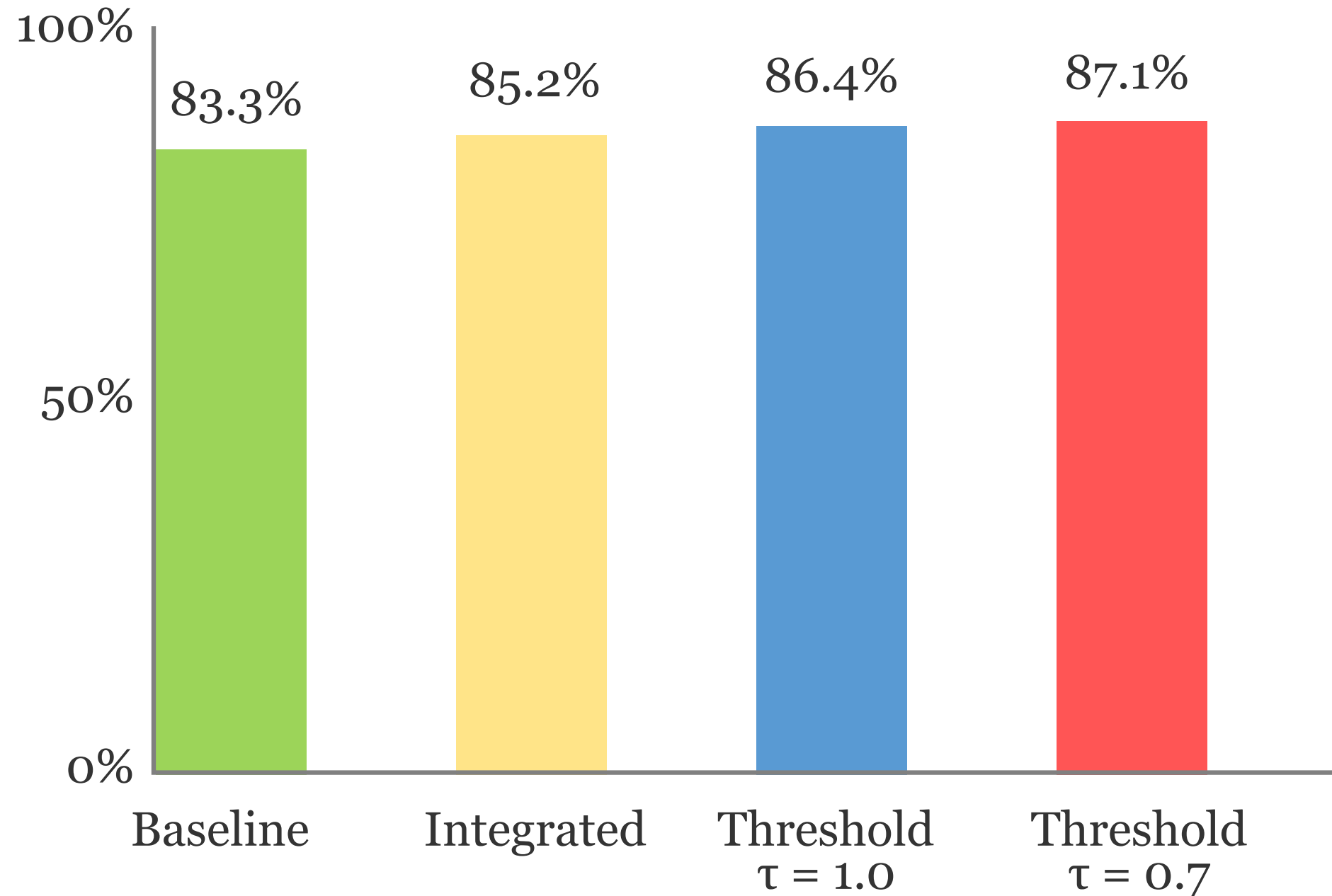
Threshold

Testing our methods

4,000 per gender

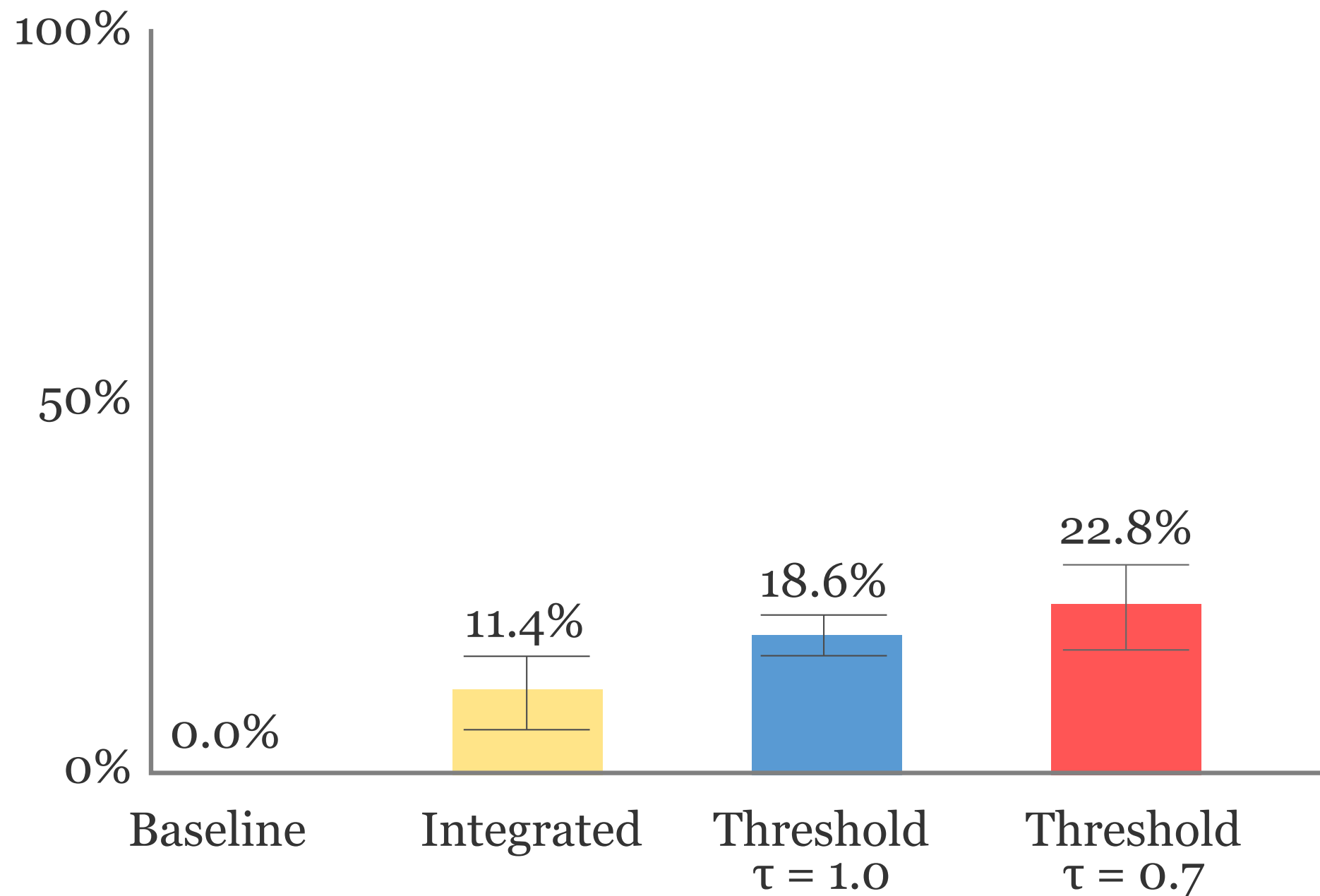
10-fold cross validation

Figure 1. SVM classifier results for all methods.



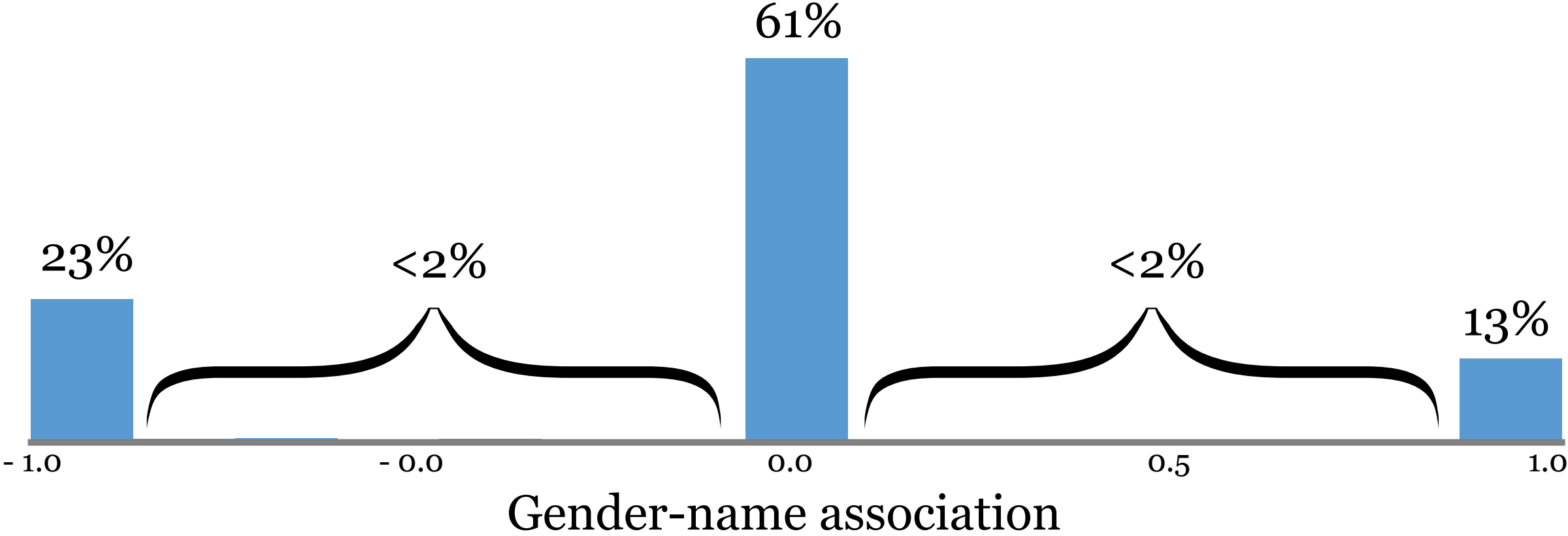
*Error bars were too small and thus were omitted

Figure 2. Improvement of each method over the baseline.



Trend: name information improves accuracy.

Figure 3. Distribution of Twitter names by gender score.



Next steps

Next steps

Improving a priori knowledge

Next steps

Improving a priori knowledge

The n-gram model

Conclusions

Conclusions

Using the name field to improve performance

Conclusions

Using the name field to improve performance

Strategy for constructing datasets

Conclusions

Using the name field to improve performance

Strategy for constructing datasets

Download our dataset: bit.ly/microtext2013

Thank you!

Wendy Liu (wendy.liu@mail.mcgill.ca) and Derek Ruths (derek.ruths@mcgill.ca)

Network Dynamics Lab (www.networkdynamics.org)

School of Computer Science, McGill University