Bleaching Text: Abstract Features for Cross-lingual Gender Prediction.

Rob van der Goot, Nikola Ljubešić, Ian Matroos,
Malvina Nissim & Barbara Plank

Bleaching Text: Abstract Features for Cross-lingual Gender Prediction.

rijksuniversiteit groningen

Jožef Stefan Institute
Ljubljana, Slovenija

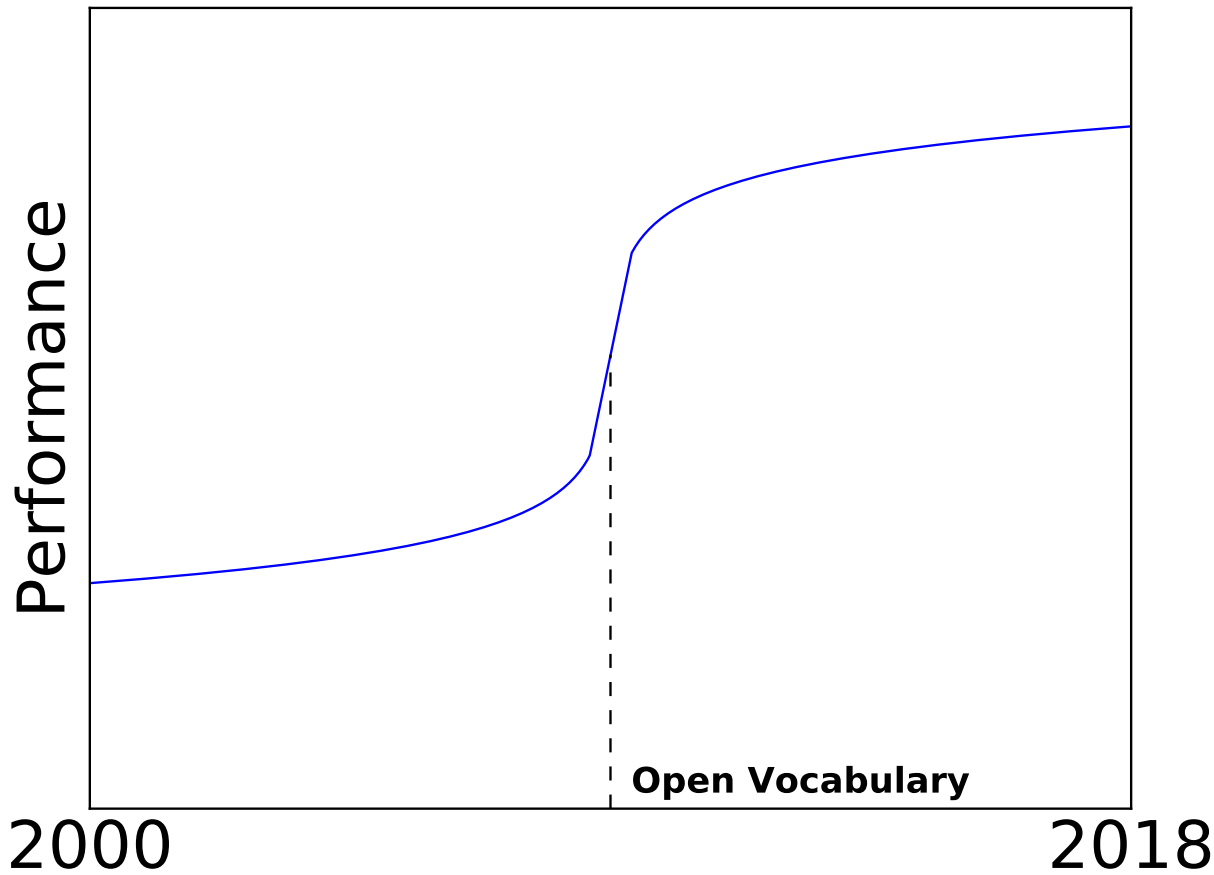
IT University of Copenhagen

Rob van der Goot, Nikola Ljubešić, Ian Matroos,
Malvina Nissim & Barbara Plank

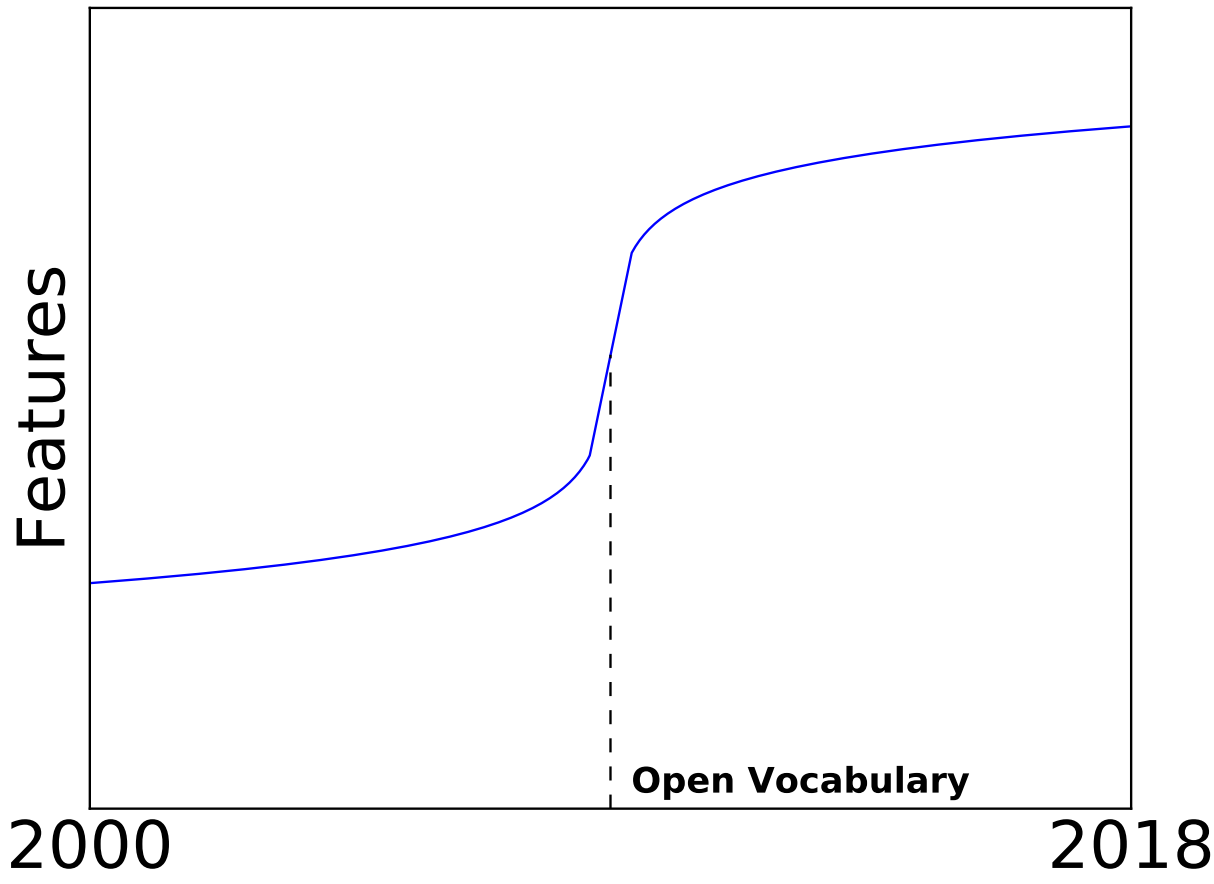
Gender Prediction

The task of predicting gender based only on text.

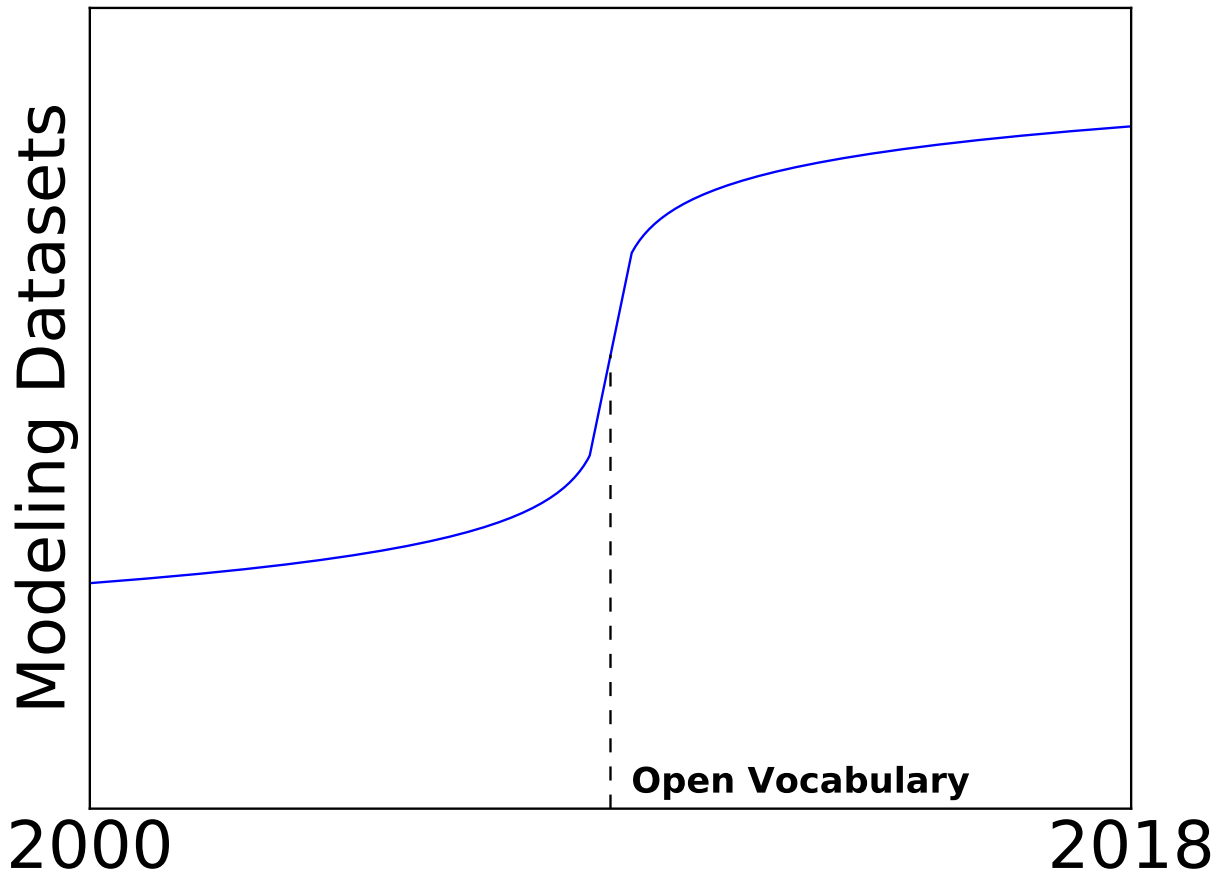
Gender Prediction



Gender Prediction



Gender Prediction



Gender Prediction

SVM with word/char n-grams performs best!

Gender Prediction

SVM with word/char n-grams performs best!

- Winner PAN 2017 shared task on author profiling:
- Words: 1-2 grams
- Characters: 3-6 grams

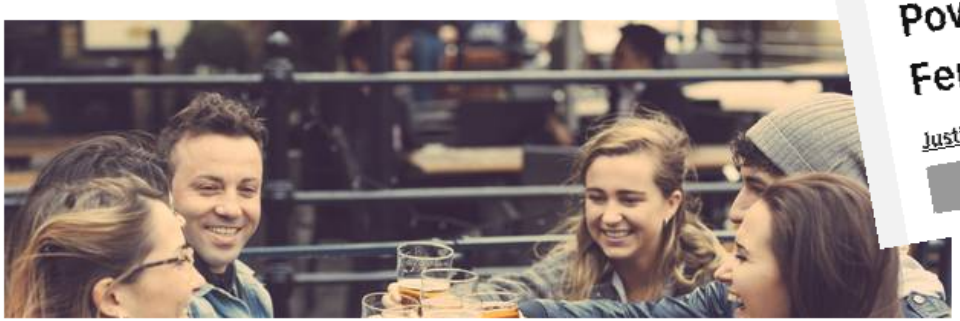
Survey: Women drinking more beer, men drinking less?

by Chris Crowell November 11, 2013

Like 1 Tweet G+

Women hop on board the beer trend

17 Feb, 2017 2:07pm



Power Hour: Craft Beer Growth Opportunity Lies with Female Consumers

Justin Kendall | Mar. 24, 2017 at 2:41 PM

https://www.brewbound.com/news/power-hour-craft-beer-growth-opportunity-lies-female-consumers
https://www.craftbrewingbusiness.com/news/survey-women-drinking-beer-men-drinking-less/
https://www.nzherald.co.nz/business/news/article.cfm?c_id=3&objectid=11802831

However, how would this lexicalized approach work across different:

- time-spans
- domains
- languages???

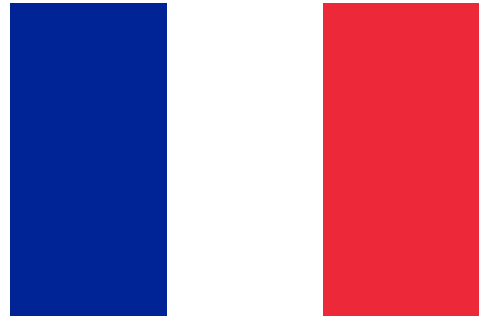
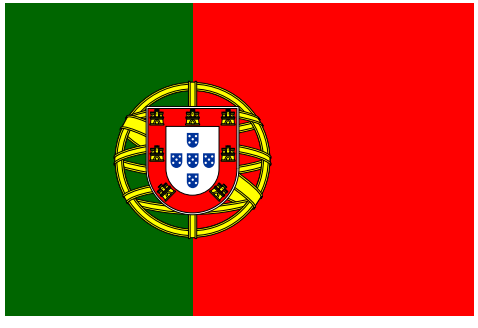
Cross-lingual Gender Prediction

- Train a model on source language(s) and evaluate on target language.

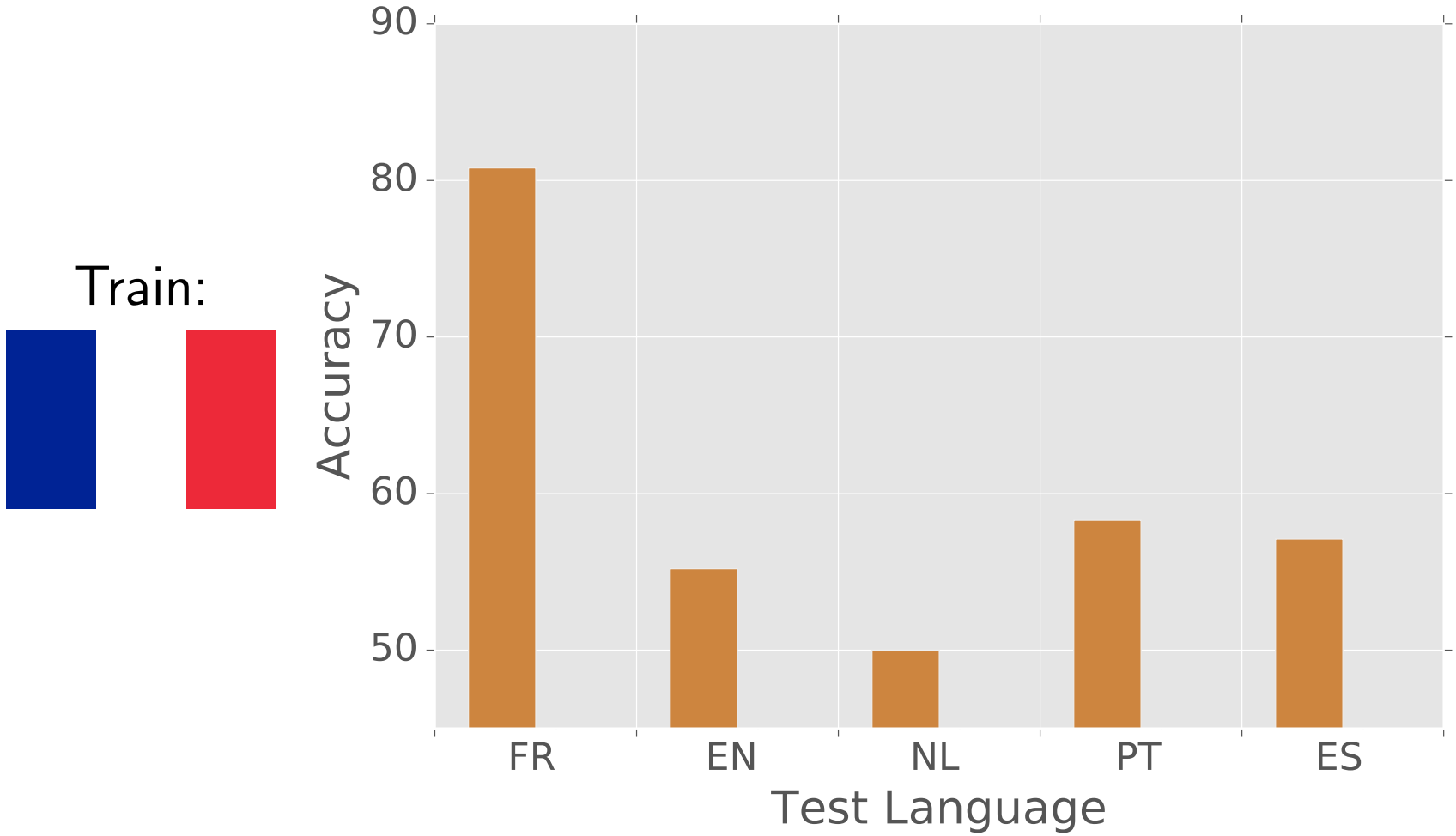
Cross-lingual Gender Prediction

- Dataset: TwiSty corpus (Verhoeven et al., 2016) + English
- 200 tweets per user, 850 - 8,112 users per language




Cross-lingual Gender Prediction



Cross-lingual Gender Prediction

USER Jaaa moeten we zeker doen ♥

Bleaching Text

Bleaching Text

Bleaching Text

Original | Massacred a bag of Doritos for lunch! 

Bleaching Text

Original	Massacred	a	bag	of	Doritos	for	lunch!	
Freq	0	5	2	5	0	5	1	0

Bleaching Text

Original	Massacred	a	bag	of	Doritos	for	lunch!	
Freq	0	5	2	5	0	5	1	0
Length	09	01	03	02	07	03	06	04

Bleaching Text

Original	Massacred	a	bag	of	Doritos	for	lunch!	
Freq	0	5	2	5	0	5	1	0
Length	09	01	03	02	07	03	06	04
PunctC	w	w	w	w	w	w	w!	

Bleaching Text

Original	Massacred	a	bag	of	Doritos	for	lunch!	
Freq	0	5	2	5	0	5	1	0
Length	09	01	03	02	07	03	06	04
PunctC	W	W	W	W	W	W	W!	
PunctA	W	W	W	W	W	W	WP	JJJJ

Bleaching Text

Original	Massacred	a	bag	of	Doritos	for	lunch!	🍆🍆🍆🍆
Freq	0	5	2	5	0	5	1	0
Length	09	01	03	02	07	03	06	04
PunctC	w	w	w	w	w	w	w!	🍆🍆🍆🍆
PunctA	w	w	w	w	w	w	wp	jjjj
Shape	ull	l	ll	ll	ull	ll	llx	xx

Bleaching Text

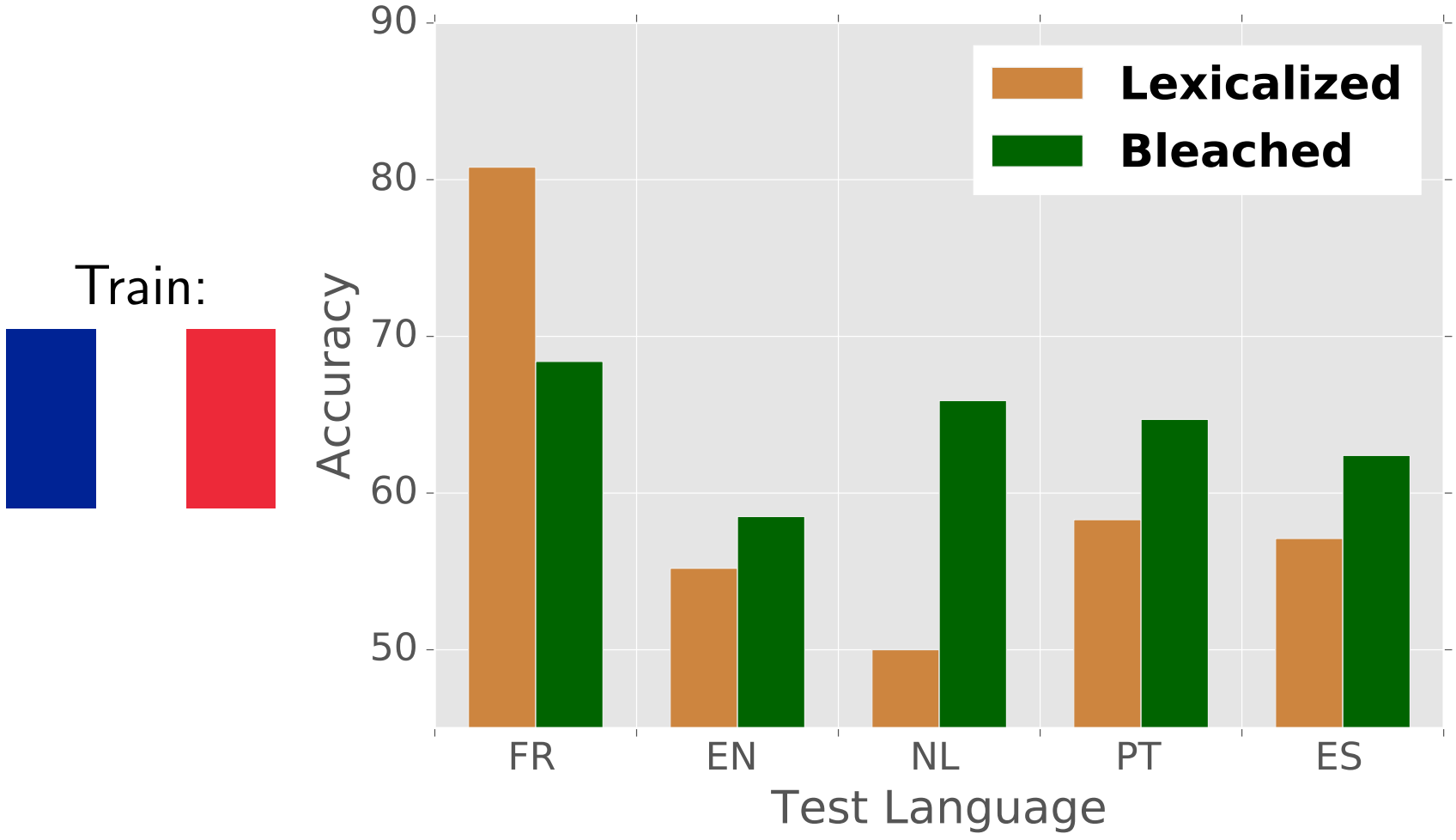
Original	Massacred	a	bag	of	Doritos	for	lunch!
Freq	0	5	2	5	0	5	1
Length	09	01	03	02	07	03	06
PunctC	w	w	w	w	w	w	w!
PunctA	w	w	w	w	w	w	wp
Shape	ull	l	ll	ll	ull	ll	llx
Vowels	cvccvccvc	v	cvc	vc	cvcvcvc	cvc	cvccco

Bleaching Text

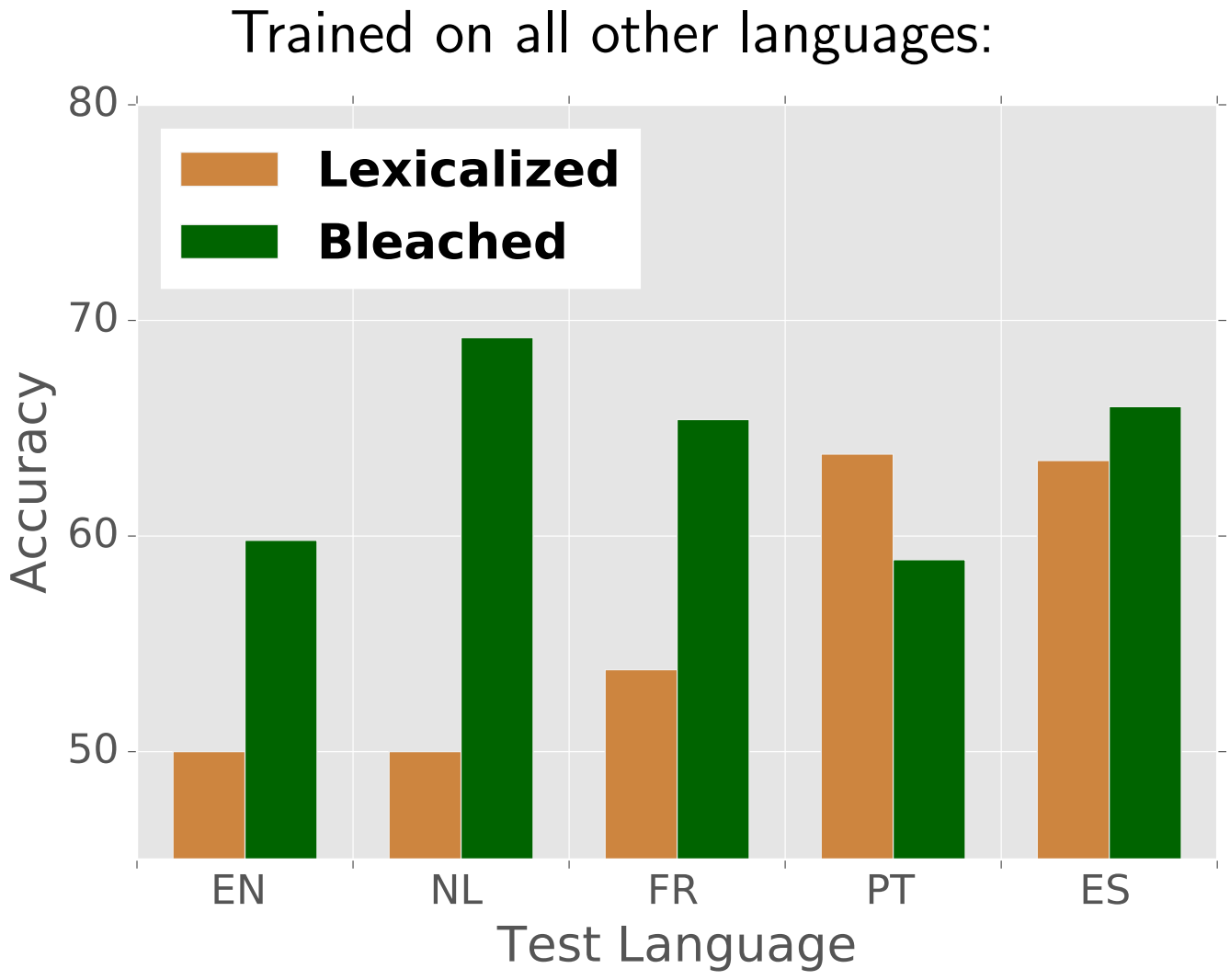
- No tokenization
- Replace usernames and URLs
- Use concatenation of the bleached representations
- Tuned in-language
- 5-grams perform best

Bleaching Text

Train:



Bleaching Text



Bleaching Text

Most predictive features

	Male	Female
1	W W W W "W"	USER E W W W
2	W W W W ?	3 5 1 5 2
3	2 5 0 5 2	W W W W ♥
4	5 4 4 5 4	E W W W W
5	W W, W W W?	LL LL LL LL LX
6	4 4 2 1 4	LL LL LL LL LUU
7	PP W W W W	W W W W *-*
8	5 5 2 2 5	W W W W JJJ
9	02 02 05 02 06	W W W W &W;W
10	5 0 5 5 2	J W W W W

Human Experiments

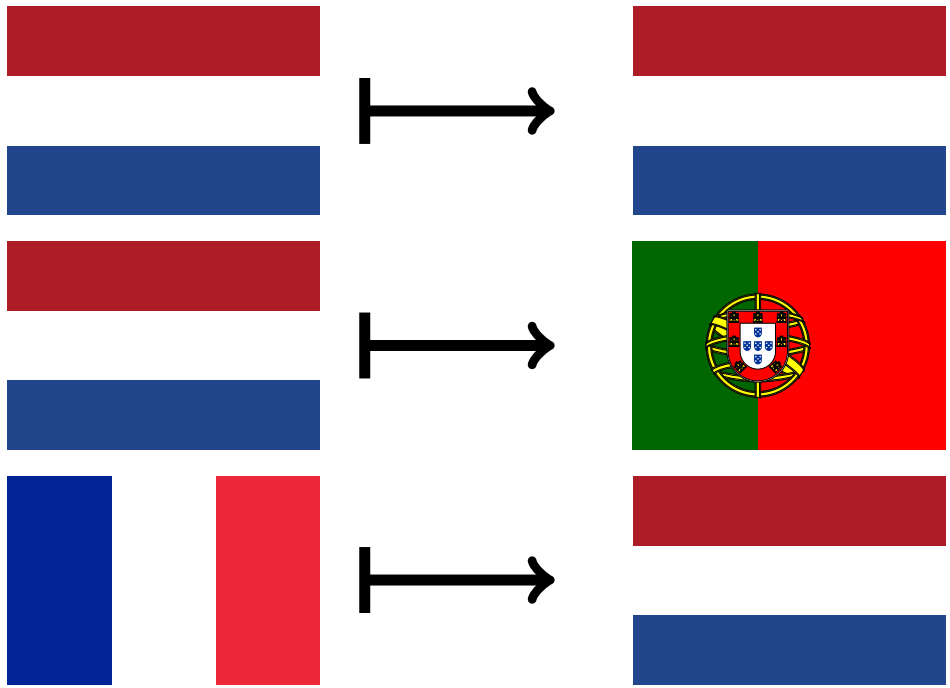
- Are humans able to predict gender based only on text for unknown languages?

Human Experiments

- 20 tweets per user (instead of 200)
- 6 annotators per language pair
- Each annotating 100 users
- 200 users per language pair, so 3 predictions per user

Human Experiments

- 20 tweets per user (instead of 200)
- 6 annotators per language pair
- Each annotating 100 users
- 200 users per language pair, so 3 predictions per user



Human Experiments

A user has posted the following tweets:

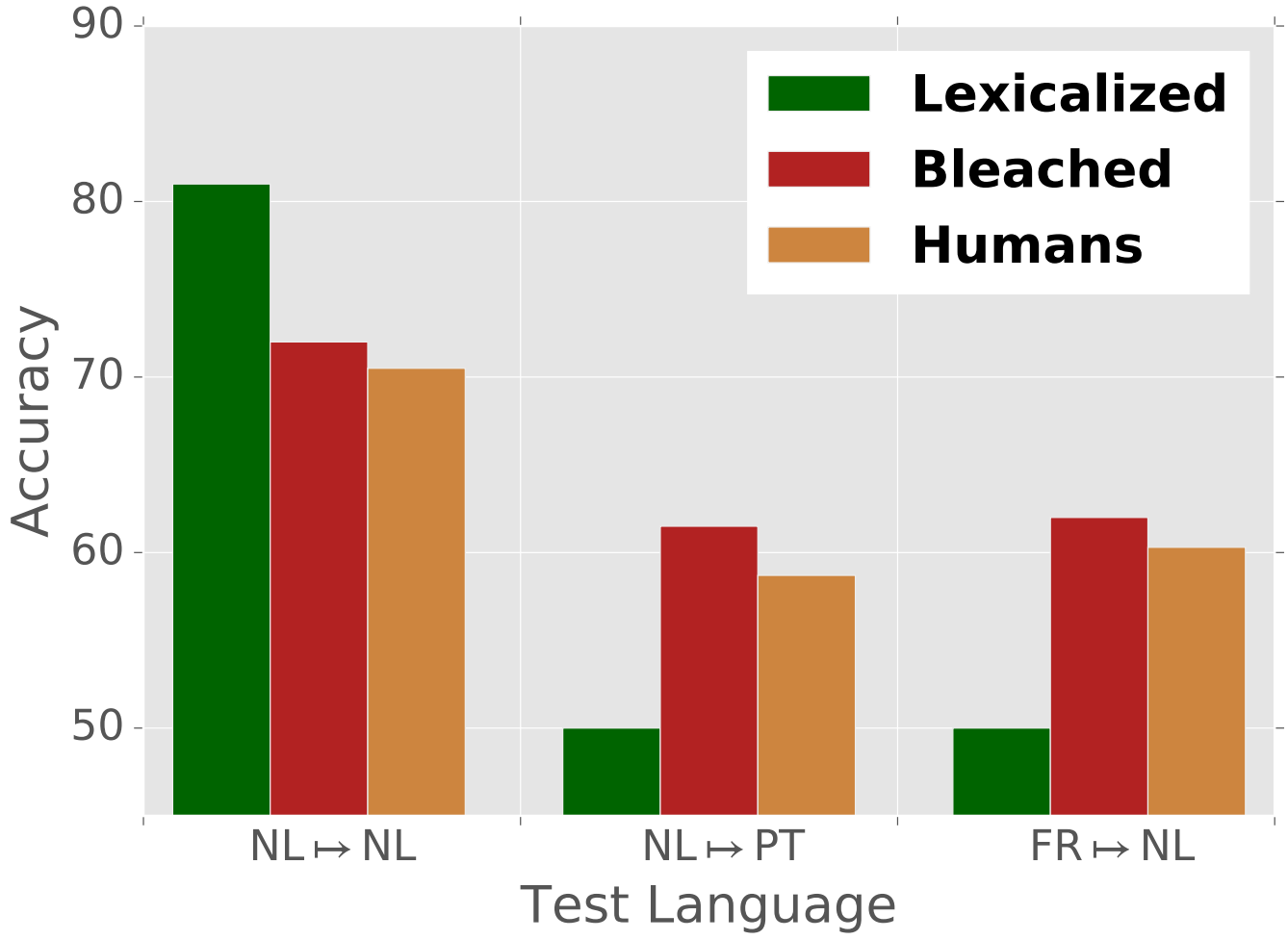
- pelo amor de deus cai na realidade URL
- a versão de REALiTi do album é tão ruim ne eu to até meio assim
- meu rosto tinha tdo pra ser ok mas nao eu tive que nascer com esse nariz horroroso e esses olhos cagados
- eu nunca ouvi nada tão lindo URL
- o mundo precisa ouvir isso URL
- USER GENTE??????????? eu apenas conciliei elas com a situação atual dá minha vida e já to todo em choque aqui pq to bateu
- meu deus eu desci o nível da timeline dum jeito q a gente já se encontra no pré sal moral
- quando a pessoa é tão medíocre que te chama de nerd debochando pq vc disse que gosta de ler
- USER bom.....eu num sei de nda
- USER eu to com o olho chei de agua sua mãe eh tão linda ♥♥♥♥
- eu definitivamente não aguento mais URL
- Rindo Muito De Meu Próprio Tweet
- USER USER sempre contribuindo para a arte de minhas amigas
- que saudade de camiliquia
- USER as arvores da minha casa tinham 70 anos.....cortaram >todas< por causa dos canos do vizinho
- USER o suprassumo da diferentona
- A NÃO paguei a lingua, pin é a terceira melhor musica do album, que musica maravilhosa
- USER USER USER qual a intenção em cmpartilhar fotos explicitas de crianças sendo abusadas?
- a minha mãe reclama de absolutamente tudo ela não para de reclamar 1 segundo, ela nunca ta de bom humor, ela nunca acha
- USER melissa do céu como assim explica

Do you think that the poster of these tweets is male or female? (required)

- ○ Male
- ○ Female

Please use your intuition.

Human Experiments

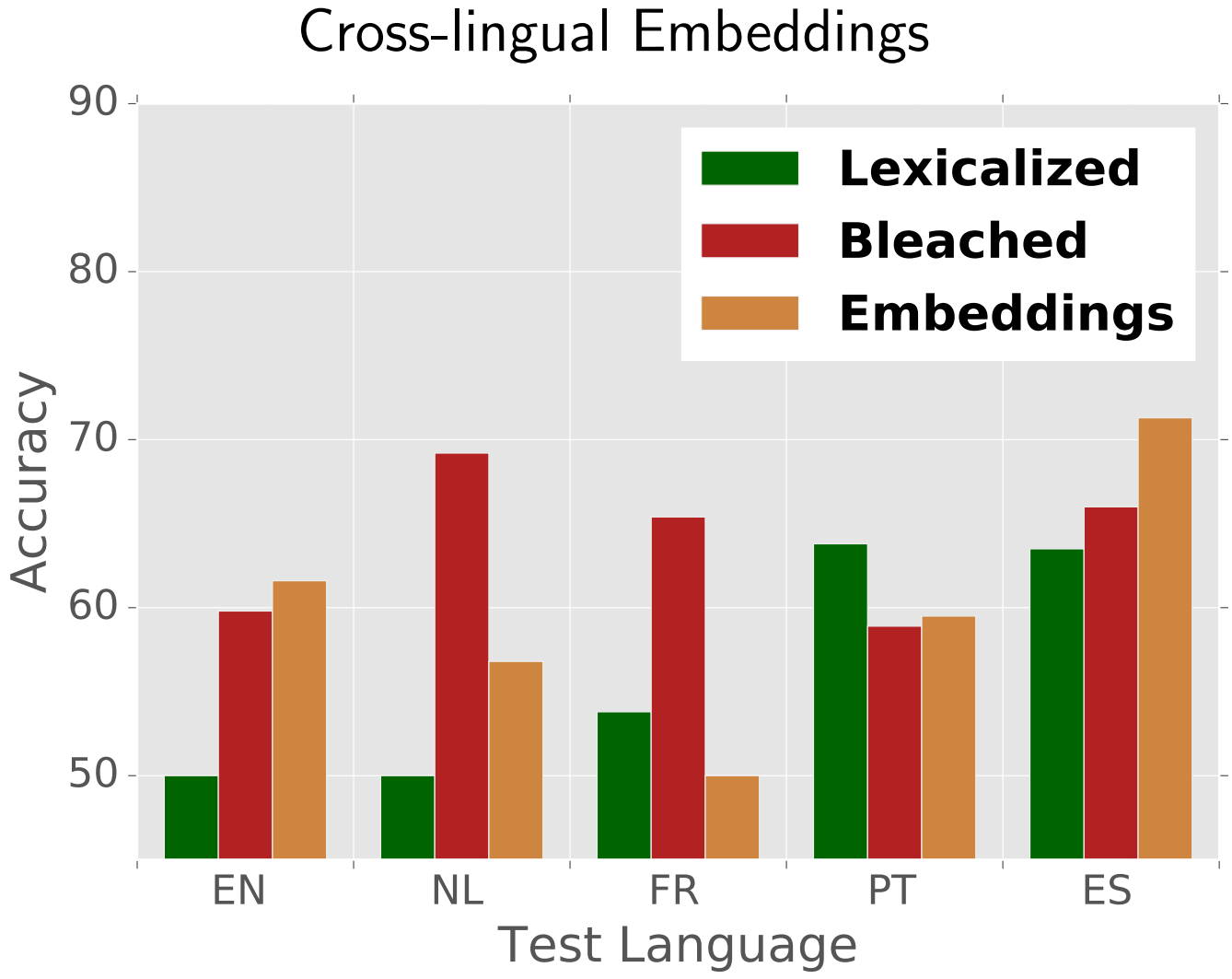


(note that the classifier had acces to 200 tweets)

Conclusions

- Lexical models break down when used cross-language
- Bleaching text improves cross-lingual performance
- Humans performance is on par with our bleached approach

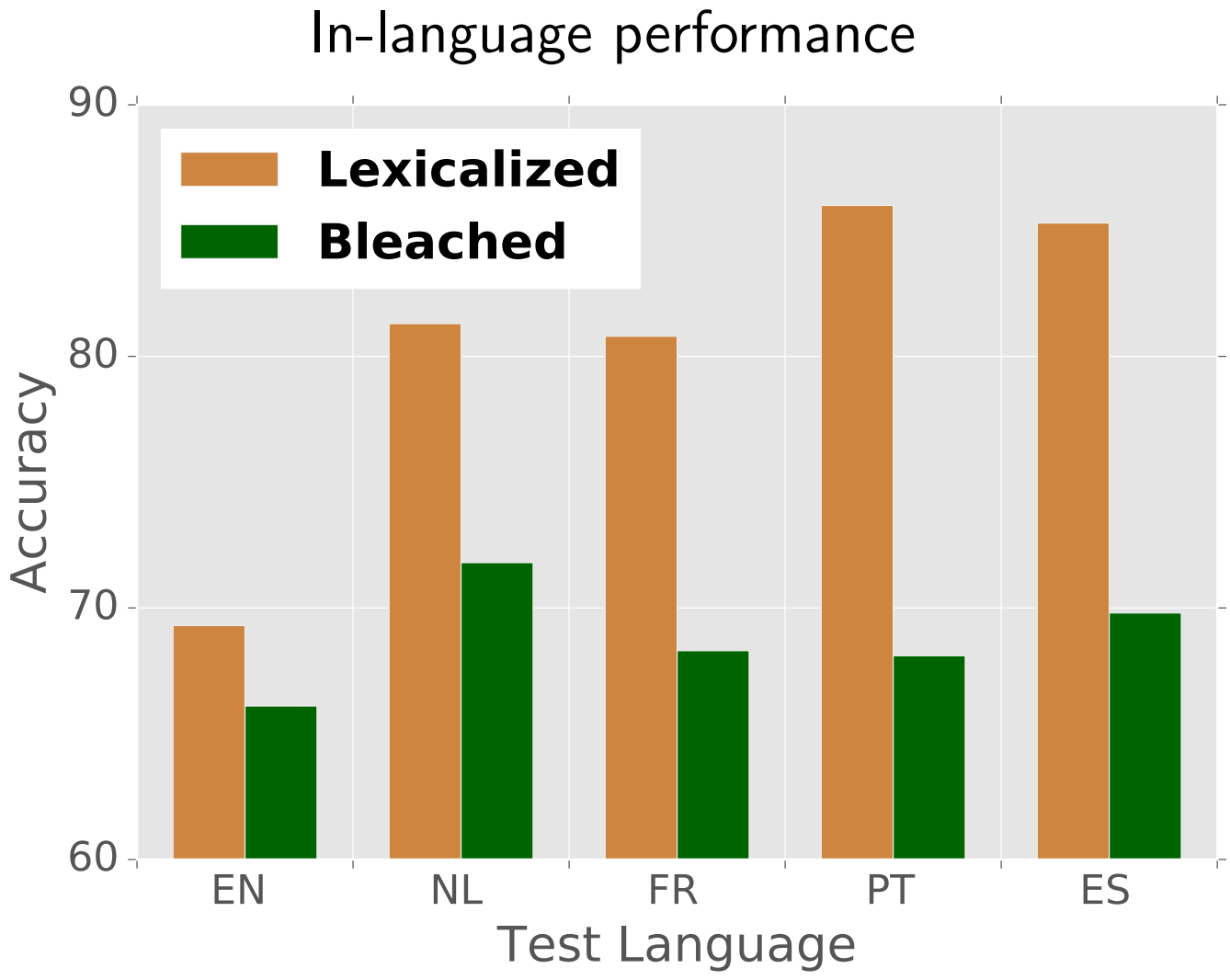
Thanks for your attention

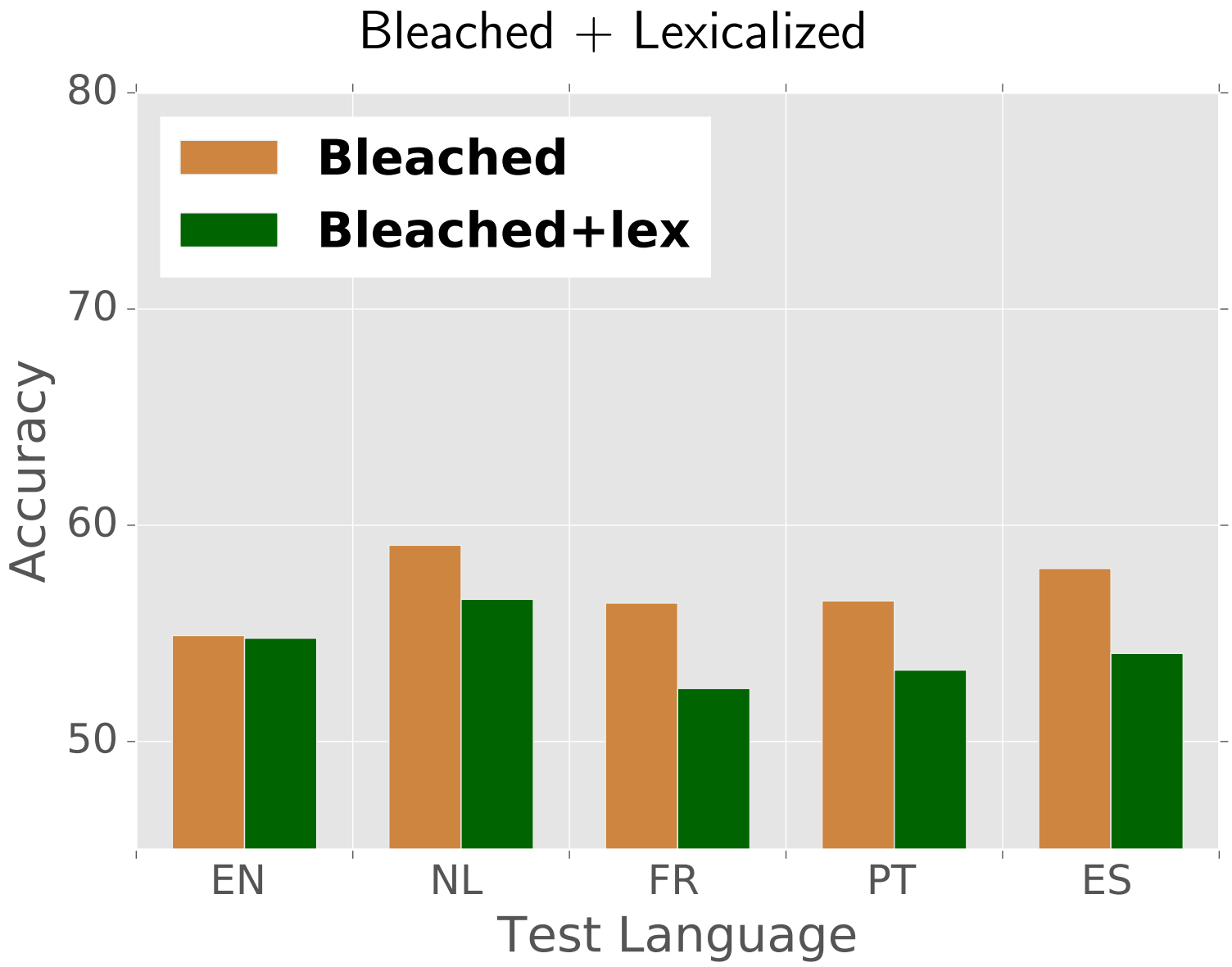


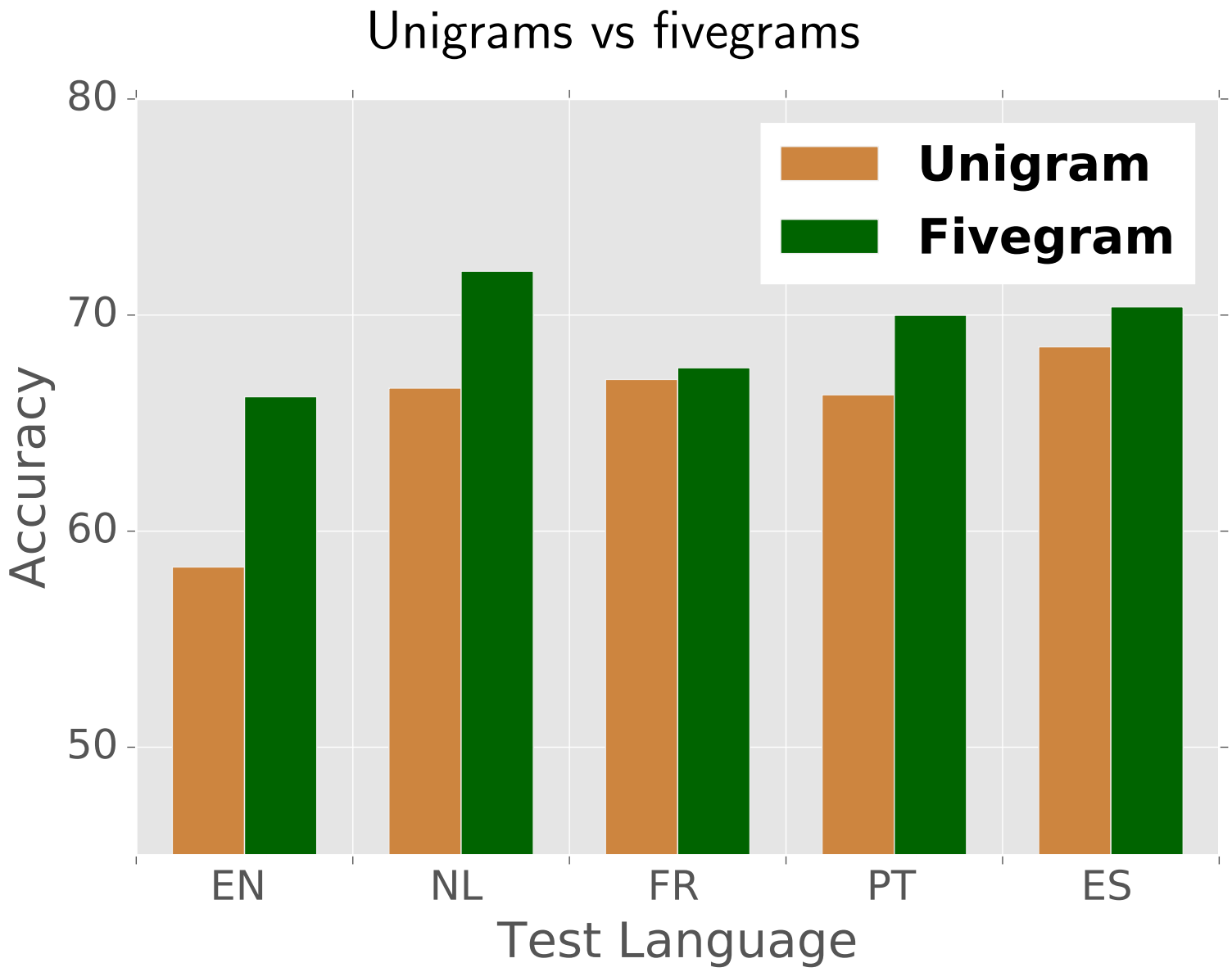
See: Plank (2017) & Smith et al. (2017)

Lexicalized Cross-language

	Test $\rightarrow$	EN	NL	FR	PT	ES
Train	EN		52.8	48.0	51.6	50.4
	NL	51.1		50.3	50.0	50.2
	FR	55.2	50.0		58.3	57.1
	PT	50.2	56.4	59.6		64.8
	ES	50.8	50.1	55.6	61.2	
	Avg	51.8	52.3	53.4	55.3	55.6

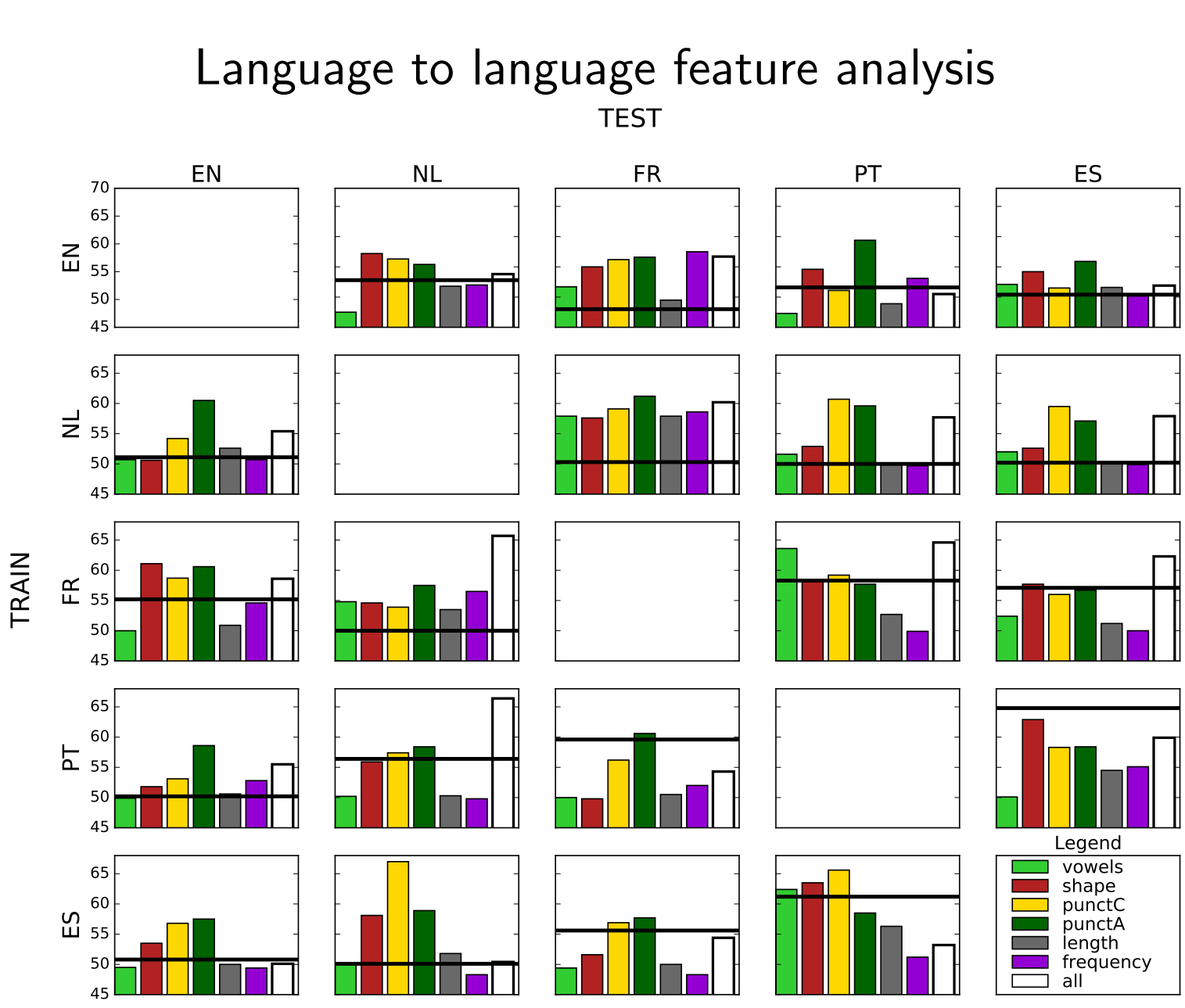

In-language performance
Lexicalized
Bleached
Accuracy
90
80
70
60
EN
NL
FR
PT
ES
Test Language


Bleached + Lexicalized
Bleached
Bleached+lex
80
70
60
50
Accuracy
EN
NL
FR
PT
ES
Test Language


Unigrams vs fivegrams
80
70
60
50
Accuracy
Unigram
Fivegram
EN
NL
FR
PT
ES
Test Language

Number of unique unigrams for Dutch

Feature	Size
Lexicalized	281011
Bleached	54103
Frequency	8
Length	79
PunctAgr	107
PunctCons	5192
Shape	2535
Vowels	46198


Language to language feature analysis
TEST
TRAIN
EN
NL
FR
PT
ES
70
65
60
55
50
45
Legend
vowels
shape
punctC
punctA
length
frequency
all