Too Good to be True: Current Approaches to Author Profiling

Malvina Nissim
Centre for Language and Cognition
University of Groningen
m.nissim@rug.nl

Language Technologies & Digital Humanities Conference
Ljubljana, 20 September 2018

Those lines that I before have writ do lie,
Even those that said I could not love you dearer:
Yet then my judgment knew no reason why
My most full flame should afterwards burn clearer.
But reckoning Time, whose million'd accidents
Creep in 'twixt vows, and change decrees of kings,
Tan sacred beauty, blunt the sharp'st intents,
Divert strong minds to the course of altering things;
Alas! why, fearing of Time's tyranny,
Might I not then say, 'Now I love you best,'
When I was certain o'er incertainty,
Crowning the present, doubting of the rest?
   Love is a babe, then might I not say so,
   To give full growth to that which still doth grow?

Why the HECK did you promise such a warm night,
But made me drive without my jacket on??? :((
To let base clouds overtake me in my way,
Hiding your bravery in their rotten smoke.
It's not enough that through the cloud you break,
To dry the rain on my storm-beaten face,
For no man well of such a salve can speak,
That heals the wound, and cures not the disgrace:
Nor can your shame give relief to my grief;
Though thou repent, yet I have still the loss:
The offender's sorrow lends but weak relief
To him that bears the strong offence's cross.
   Ah! but those tears are pearl which your love sheds,
   And they are rich and ransom all ill deeds.

When rampant rumor doth my ears confound
And insult hound me for my mere shape’s sake,
Then do I pause to sit upon the ground
And tell sad stories of the perjured snake.
The worthy serpent by the world full curst
In truth is innocent and full of grace;
His dimensions all compact, his mind well versed.
Why therefore villain? Wherefore base?
Regard my supple body lithe and thin,
My curling arabesques, my twin-tounged kiss;
Hath not a serpent flesh, bones, skin?
If struck, nay, trod upon, shall he not hiss?
   Fie, hedge-pigs. Ye are slanderers most vile;
   Unworthy e’en to speak the name, Reptile.

I grew up watching many war films. @HacksawRidge is the best one I've seen. Had me on edge, feels like you're there!

I grew up watching many war films. @HacksawRidge is the best one I've seen. Had me on edge, feels like you're there!

F

Division 1 Champions!! Well done King @JakeButler17 #champions #waitakerecity #prmierleaguebound #kingbulter

Division 1 Champions!! Well done King @JakeButler17 #champions #waitakerecity #prmierleaguebound #kingbulter

M

Welcomed a new addition to the family today. So cute! https://t.co/M2x06UzKQx

Welcomed a new addition to the family today. So cute! https://t.co/M2x06UzKQx

M

@_tahliaa your wellbeing is always a priority, please make sure you're taking care of yourself 💕

@_tahliaa your wellbeing is always a priority, please make sure you're taking care of yourself 💕

F

The build quality of the #nintendoswitch seems... cheap. But still. 😍😍😍

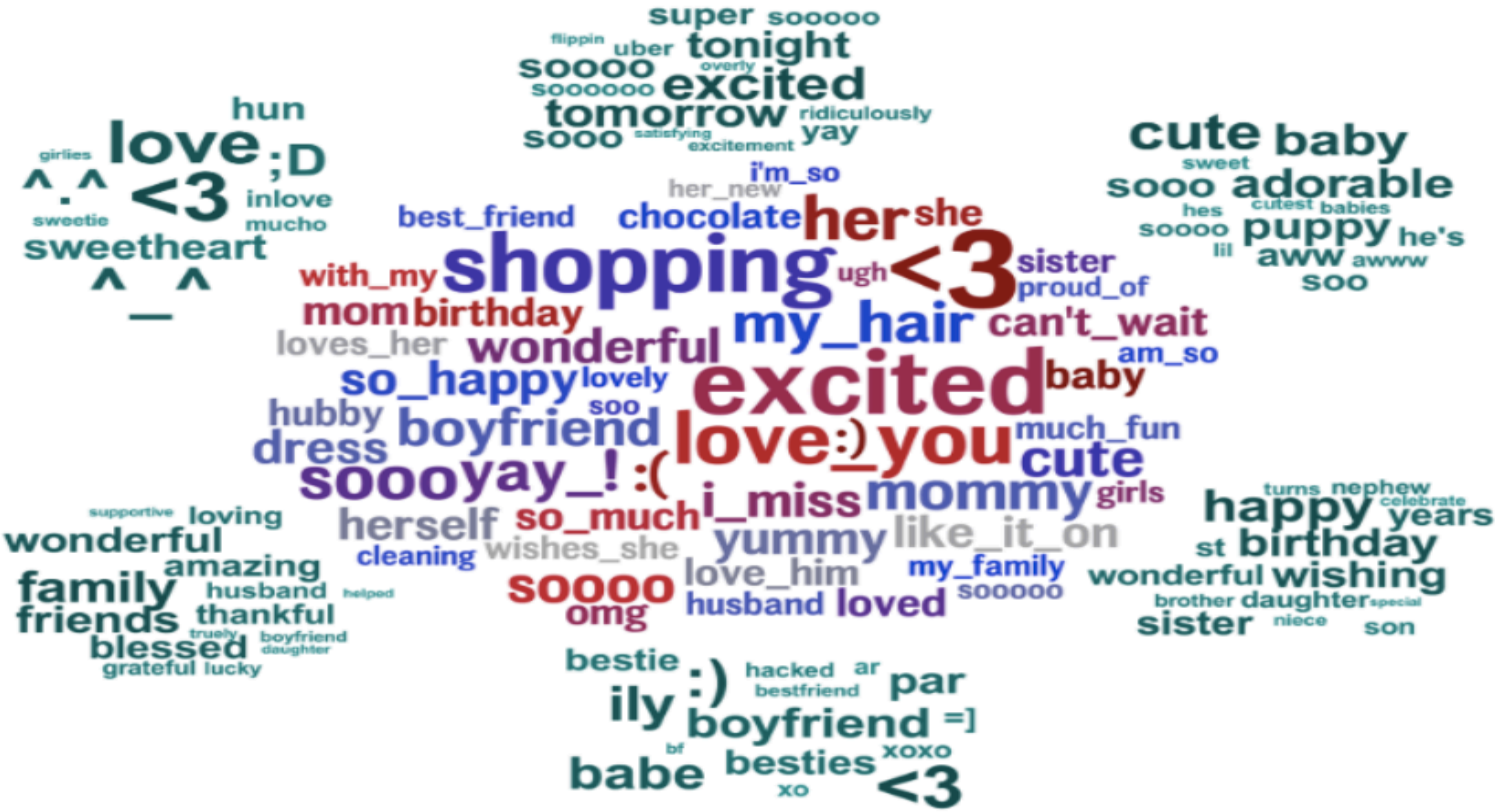
The build quality of the #nintendoswitch seems... cheap. But still. 😍😍😍

M

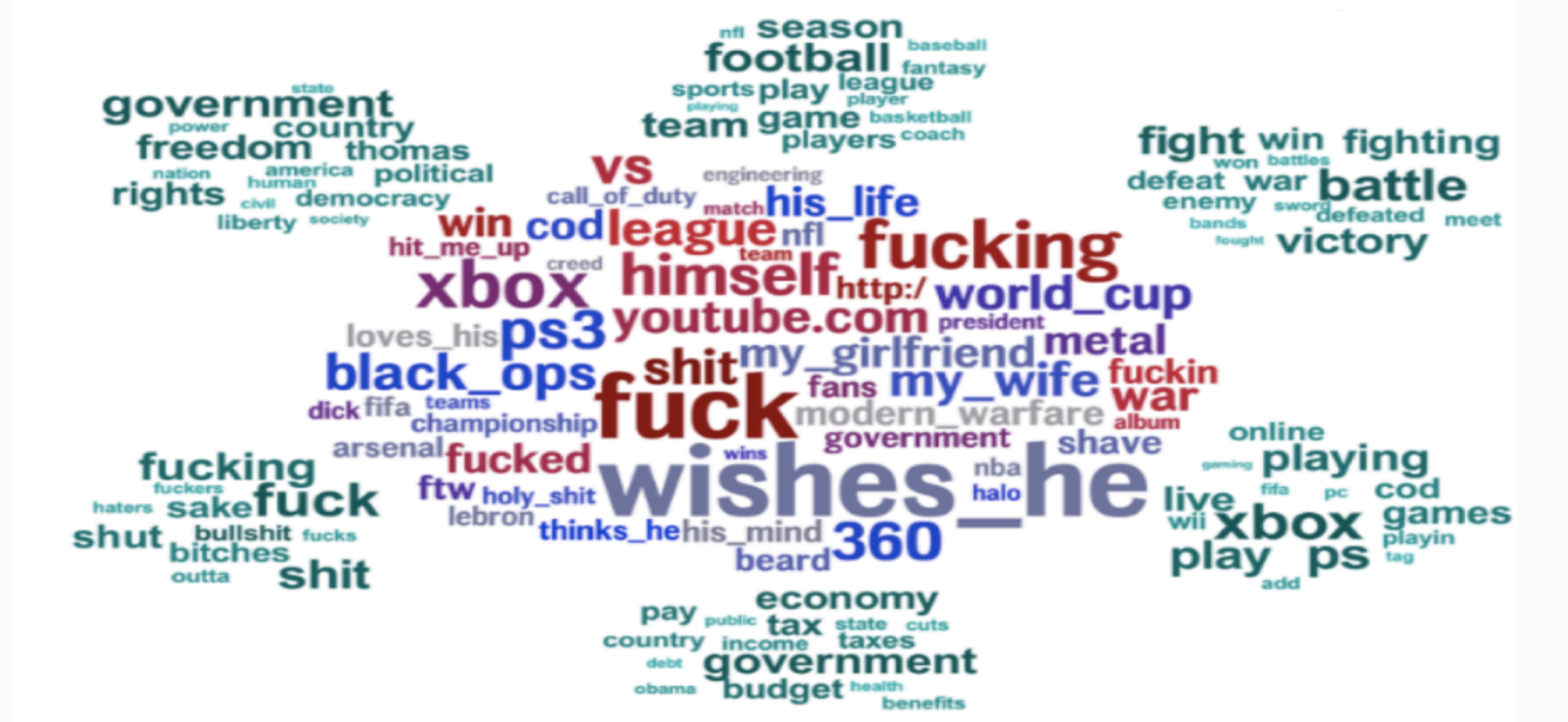
“[humans] relied on correct stereotypes, but relied on them more heavily than warranted by data.

For example, [they] assume that males post more than they do about sports and business, females show more joy, older users more interest in politics and younger users use more slang and are more self-referential.”

Lucie Flekova et al. (2016). “Analysing Biases in Human Perception of User Age and Gender from Text”, Proceedings of ACL 2016.

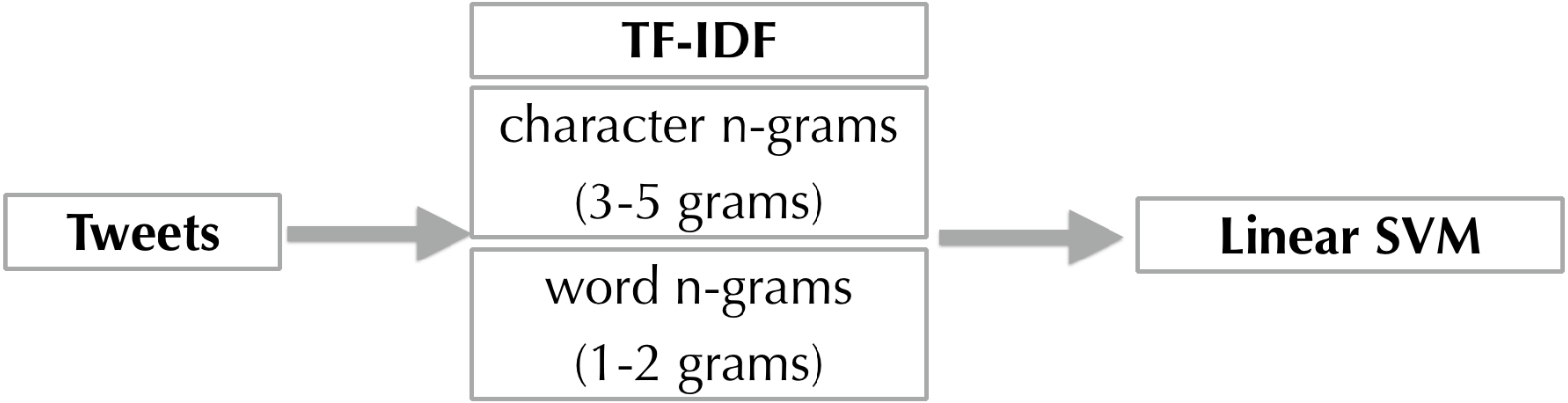


(source: Andrew Schwartz et al. (2013), Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PlosONE)



(source: Andrew Schwartz et al. (2013), Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PlosONE)

N-gram-based SVM

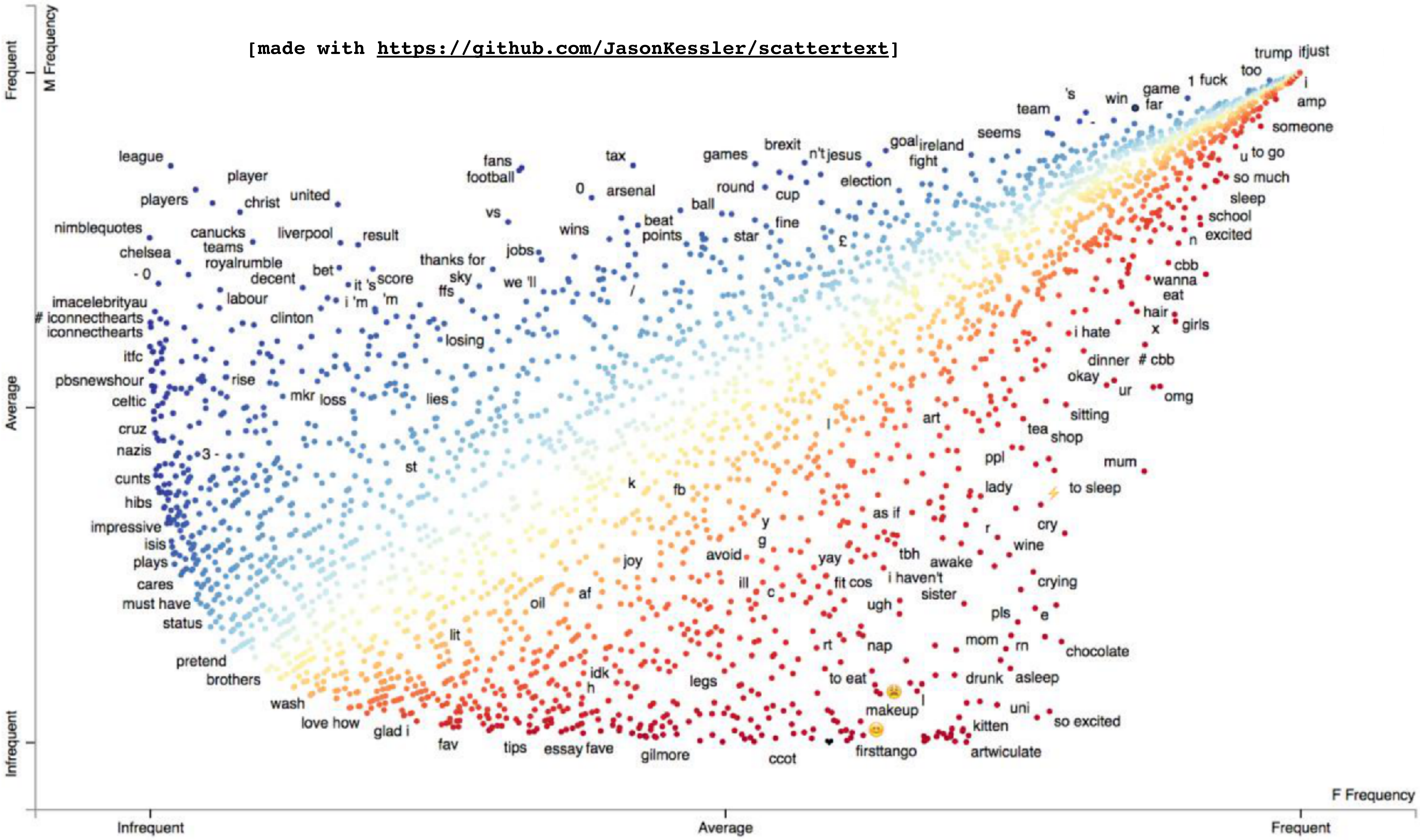

Tweets

TF-IDF

character n-grams (3-5 grams)

word n-grams (1-2 grams)

Linear SVM

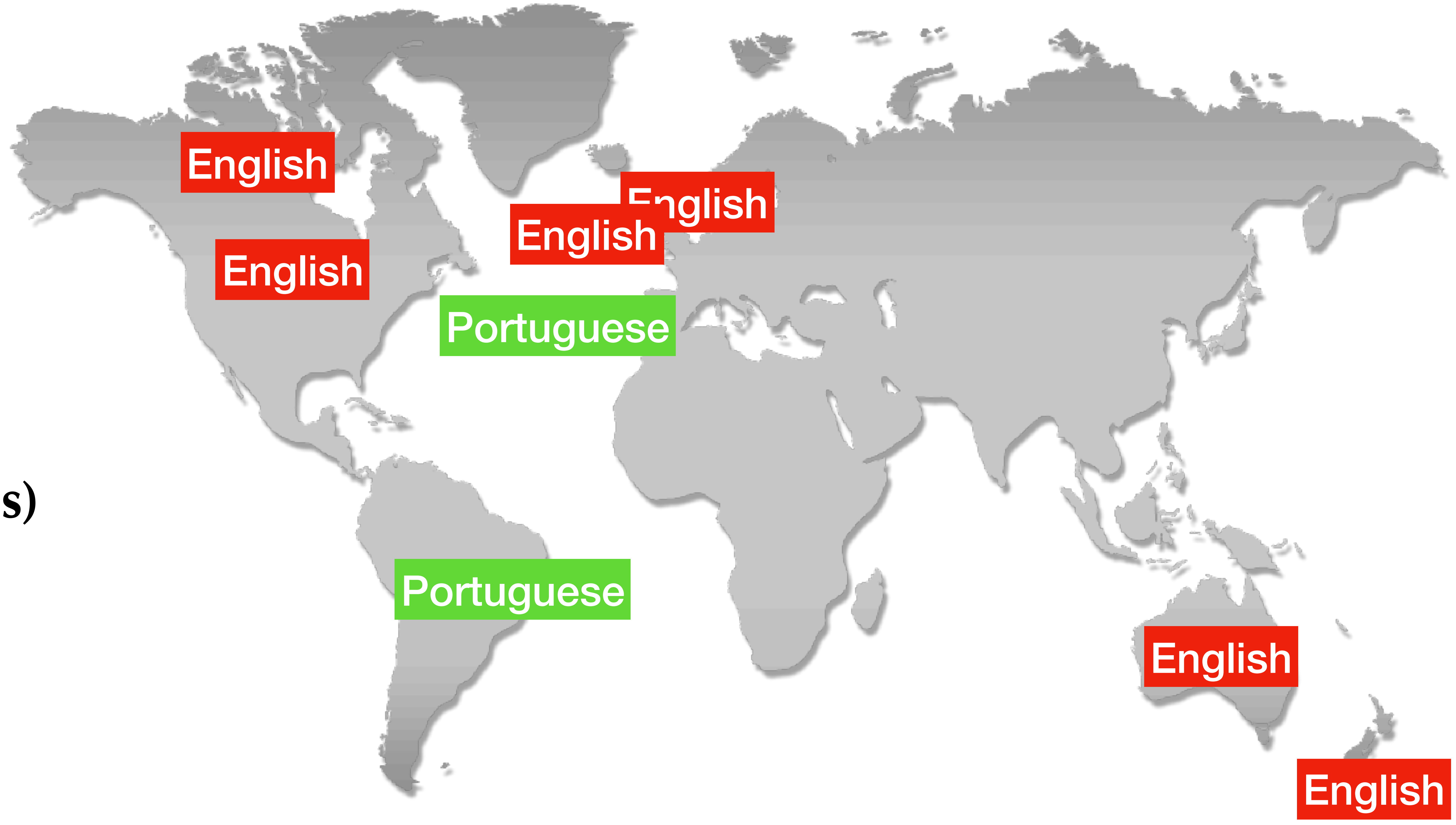


[made with https://github.com/JasonKessler/scattertext]

PAN 2017 challenge on author profiling

Gender detection

Language Variety detection (four languages)



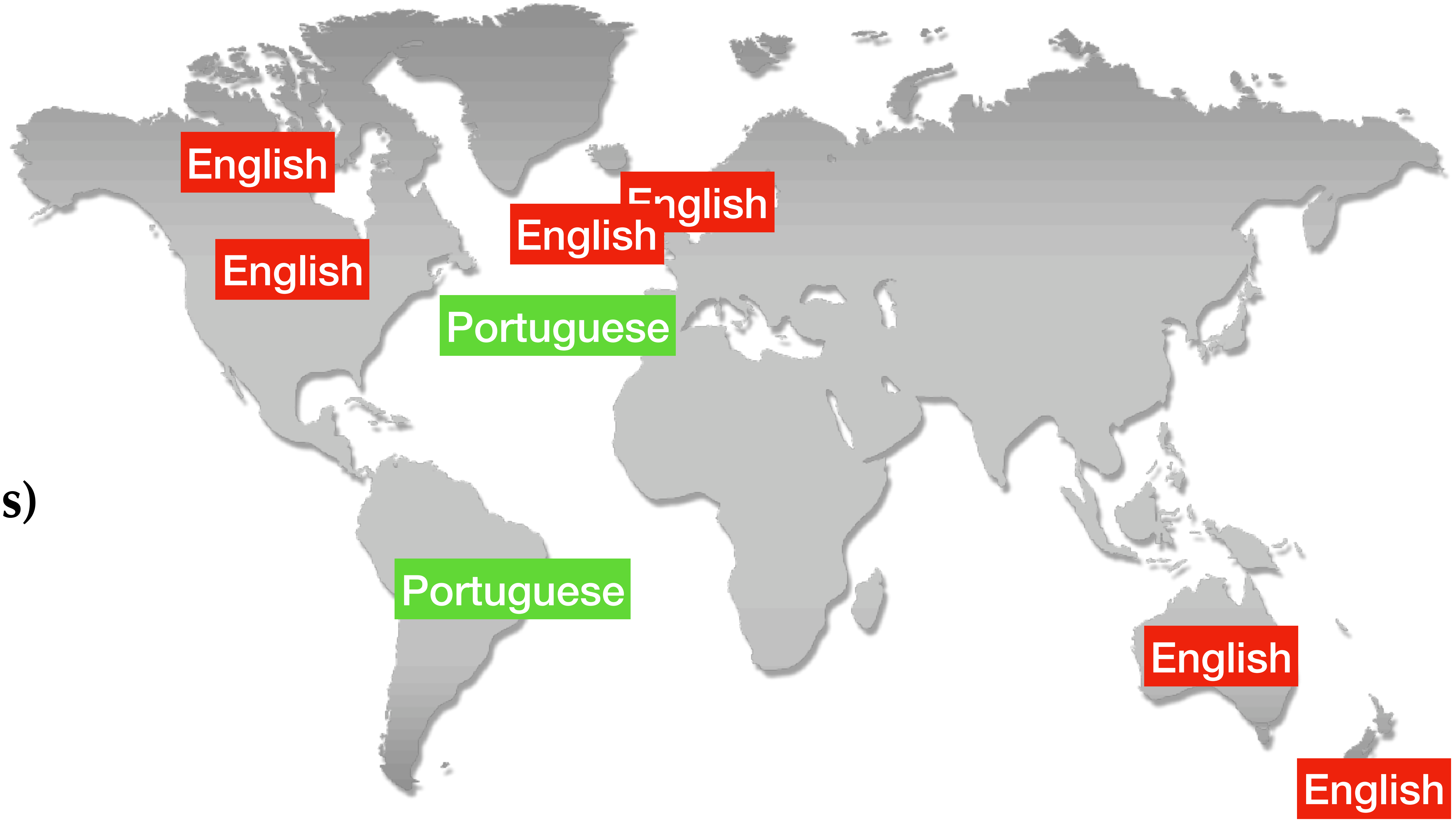
- **English** (Australia, Canada, Great Britain, Ireland, New Zealand, United States)
- **Spanish** (Argentina, Chile, Colombia, Mexico, Peru, Spain, Venezuela)
- **Portuguese** (Brazil, Portugal)
- **Arabic** (Egypt, Gulf, Levantine, Maghrebi)

PAN 2017 challenge on author profiling

Gender detection

Language Variety detection (four languages)

Language	#Varieties	#Authors
Arabic	4	2400
English	6	3600
Portuguese	2	1200
Spanish	7	4200



- ✦ ~100 tweets / author
- ✦ 600 authors / variety

There’s MORE MORE MORE

Data

Features

Classifiers

There's MORE MORE MORE

Data	Features	Classifiers
✗		

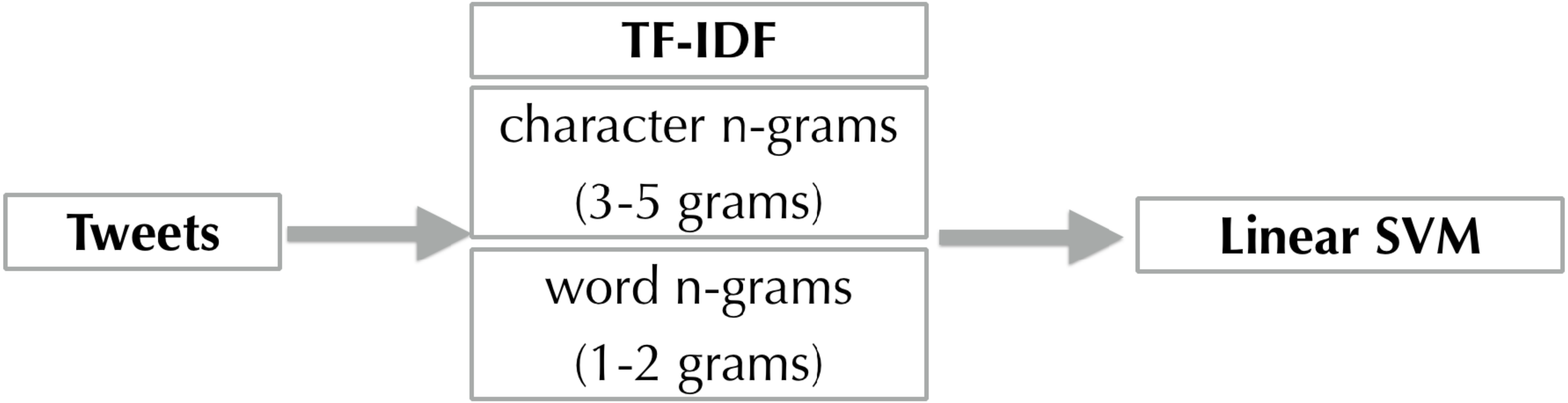
There's MORE MORE MORE

Data	Features	Classifiers
✗	✗	

There’s MORE MORE MORE

Data	Features	Classifiers
✗	✗	✗

N-gram-based SVM



Tweets

TF-IDF

character n-grams (3-5 grams)

word n-grams (1-2 grams)

Linear SVM

PAN 2017 challenge on author profiling

Task	System	Arabic	English	Portuguese	Spanish	Average	+ 2nd
Variety	N-GrAM	**0.8313**	0.8988	0.9813	0.9621	0.9184	0.0013
	LDR	0.8250	**0.8996**	**0.9875**	**0.9625**	**0.9187**	
Gender	N-GrAM	**0.8006**	**0.8233**	**0.8450**	**0.8321**	**0.8253**	0.0029
	LDR	0.7044	0.7220	0.7863	0.7171	0.7325	
Joint	N-GrAM	**0.6831**	**0.7429**	**0.8288**	**0.8036**	**0.7646**	0.0101
	LDR	0.5888	0.6357	0.7763	0.6943	0.6738	

Author Profiling ‹ 2017 ›

Authorship analysis deals with the classification of texts into classes based on the stylistic choices of their authors. Beyond the author identification and author verification tasks where the style of individual authors is examined, author profiling distinguishes between classes of authors studying their sociolect aspect, that is, how language is shared by people. This helps in identifying profiling aspects such as gender, age, native language, or personality type. Author profiling is a problem of growing importance in applications in forensics, security, and marketing. E.g., from a forensic linguistics perspective one would like being able to know the linguistic profile of the author of a harassing text message (language used by a certain type of people) and identify certain characteristics (language as evidence). Similarly, from a marketing viewpoint, companies may be interested in knowing, on the basis of the analysis of blogs and online product reviews, the demographics of people that like or dislike their products. The focus is on author profiling in social media since we are mainly interested in everyday language and how it reflects basic social and personality processes.

meaning cloud

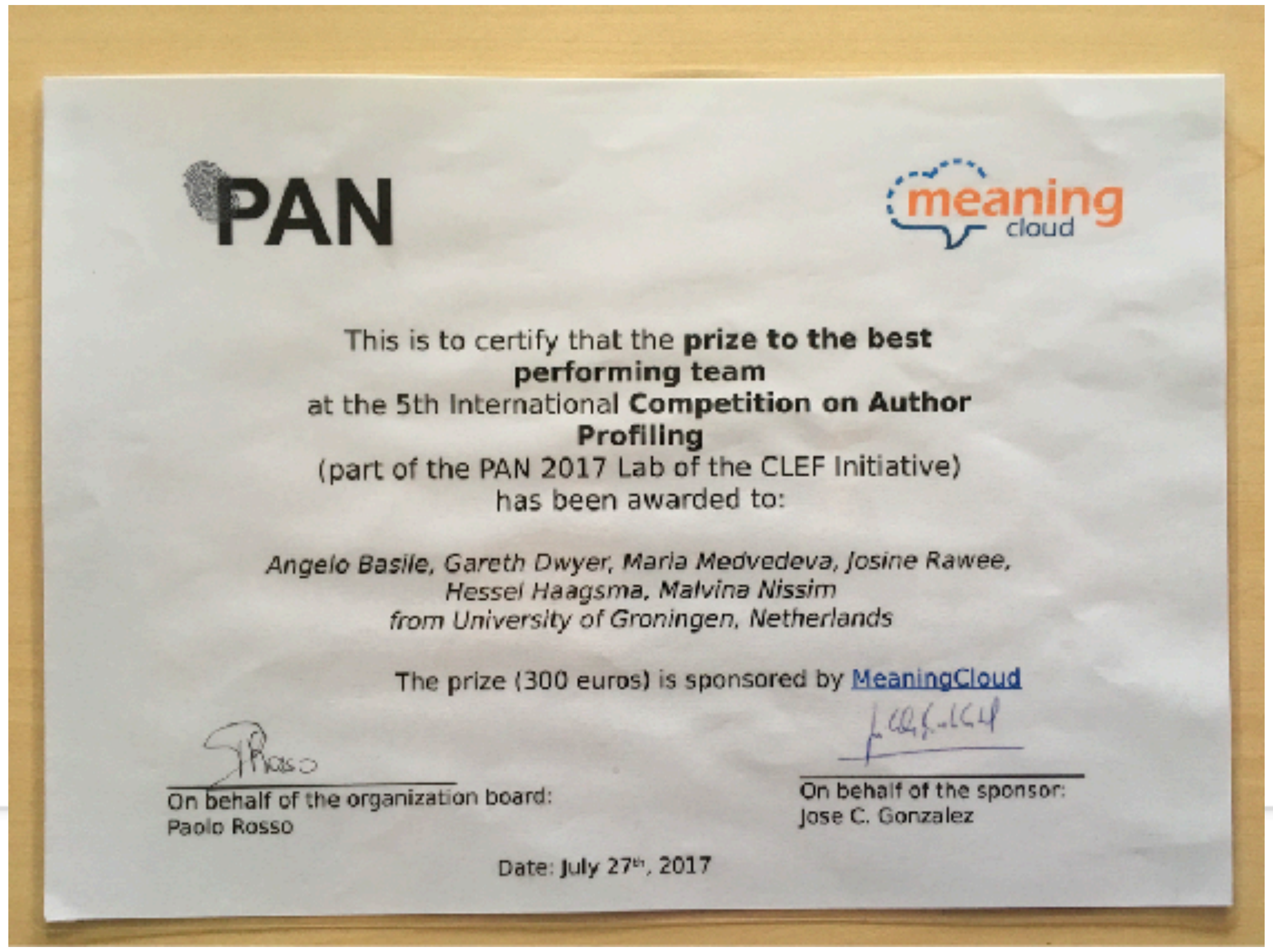
Sponsor

Award

We are happy to announce that the best performing team at the 5th International Competition on Author Profiling will be awarded 300,- Euro sponsored by MeaningCloud.

- Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. University of Groningen, Netherlands.

Congratulations!



PAN

meaning cloud

This is to certify that the **prize to the best performing team** at the 5th International **Competition on Author Profiling** (part of the PAN 2017 Lab of the CLEF Initiative) has been awarded to:

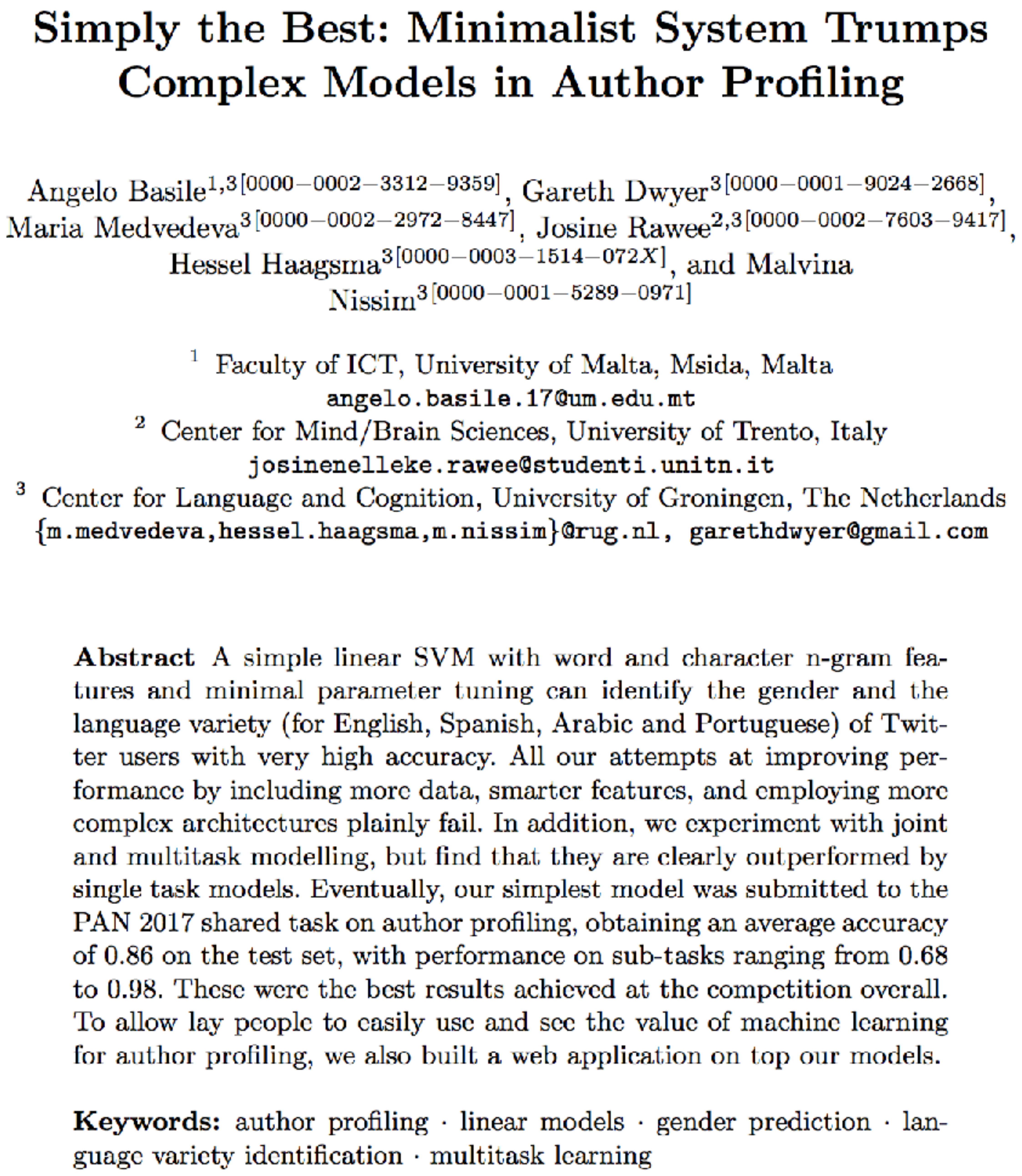
Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, Malvina Nissim from University of Groningen, Netherlands

The prize (300 euros) is sponsored by MeaningCloud

On behalf of the organization board:
Paolo Rosso

On behalf of the sponsor:
Jose C. Gonzalez

Date: July 27th, 2017



Simply the Best: Minimalist System Trumps Complex Models in Author Profiling

Angelo Basile[1,3][0000–0002–3312–9359], Gareth Dwyer[3][0000–0001–9024–2668], Maria Medvedeva[3][0000–0002–2972–8447], Josine Rawee[2,3][0000–0002–7603–9417], Hessel Haagsma[3][0000–0003–1514–072X], and Malvina Nissim[3][0000–0001–5289–0971]

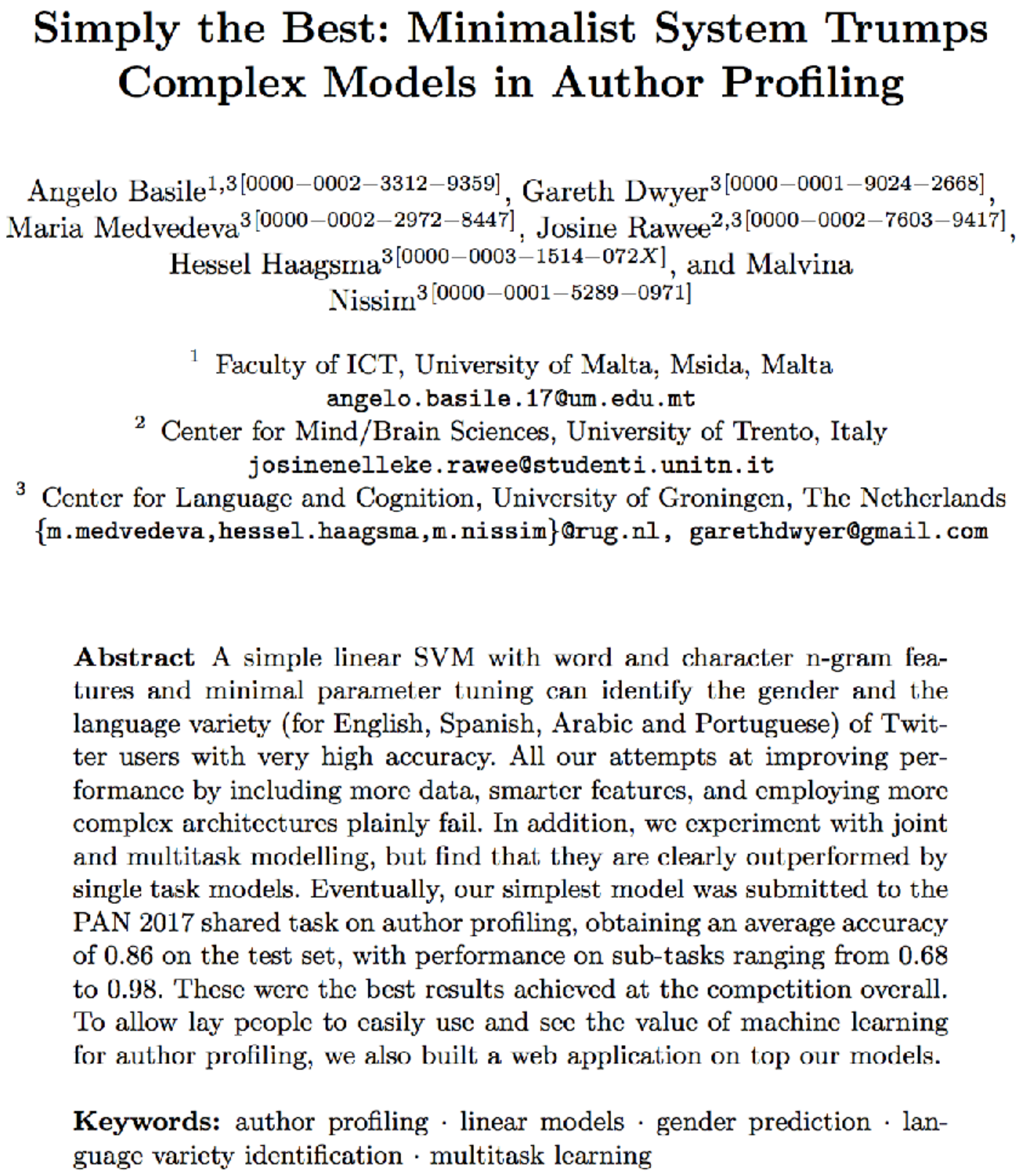
[1] Faculty of ICT, University of Malta, Msida, Malta
angelo.basile.17@um.edu.mt
[2] Center for Mind/Brain Sciences, University of Trento, Italy
josinenelleke.rawee@studenti.unitn.it
[3] Center for Language and Cognition, University of Groningen, The Netherlands
{m.medvedeva,hessel.haagsma,m.nissim}@rug.nl, garethdwyer@gmail.com

Abstract A simple linear SVM with word and character n-gram features and minimal parameter tuning can identify the gender and the language variety (for English, Spanish, Arabic and Portuguese) of Twitter users with very high accuracy. All our attempts at improving performance by including more data, smarter features, and employing more complex architectures plainly fail. In addition, we experiment with joint and multitask modelling, but find that they are clearly outperformed by single task models. Eventually, our simplest model was submitted to the PAN 2017 shared task on author profiling, obtaining an average accuracy of 0.86 on the test set, with performance on sub-tasks ranging from 0.68 to 0.98. These were the best results achieved at the competition overall. To allow lay people to easily use and see the value of machine learning for author profiling, we also built a web application on top our models.

Keywords: author profiling · linear models · gender prediction · language variety identification · multitask learning

Q1: Why do complex models fail so miserably?

Q2: Why do simple models work so well?



Simply the Best: Minimalist System Trumps Complex Models in Author Profiling

Angelo Basile[1,3][0000−0002−3312−9359], Gareth Dwyer[3][0000−0001−9024−2668], Maria Medvedeva[3][0000−0002−2972−8447], Josine Rawee[2,3][0000−0002−7603−9417], Hessel Haagsma[3][0000−0003−1514−072X], and Malvina Nissim[3][0000−0001−5289−0971]

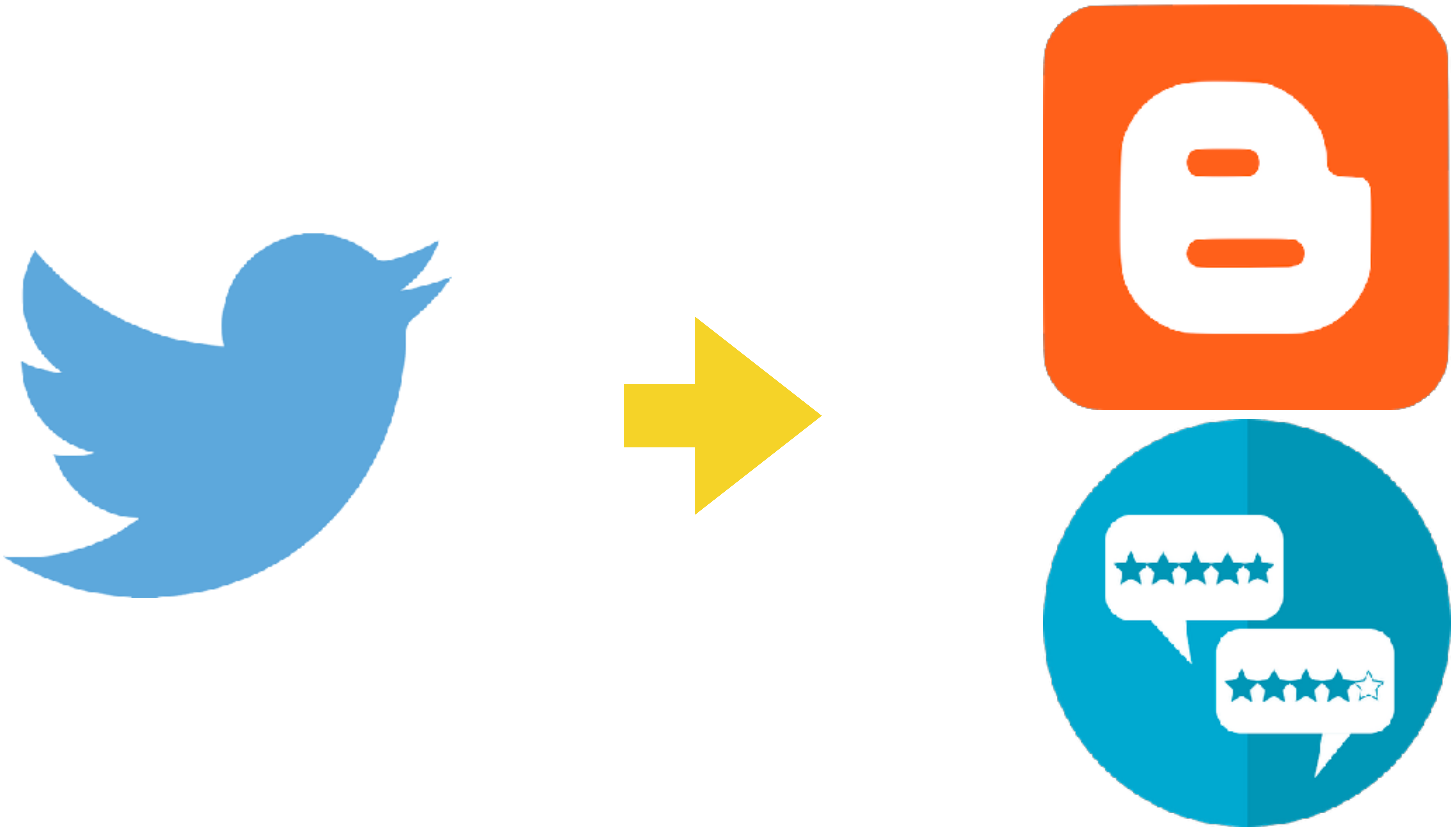
[1] Faculty of ICT, University of Malta, Msida, Malta
angelo.basile.17@um.edu.mt
[2] Center for Mind/Brain Sciences, University of Trento, Italy
josinenelleke.rawee@studenti.unitn.it
[3] Center for Language and Cognition, University of Groningen, The Netherlands
{m.medvedeva,hessel.haagsma,m.nissim}@rug.nl, garethdwyer@gmail.com

Abstract A simple linear SVM with word and character n-gram features and minimal parameter tuning can identify the gender and the language variety (for English, Spanish, Arabic and Portuguese) of Twitter users with very high accuracy. All our attempts at improving performance by including more data, smarter features, and employing more complex architectures plainly fail. In addition, we experiment with joint and multitask modelling, but find that they are clearly outperformed by single task models. Eventually, our simplest model was submitted to the PAN 2017 shared task on author profiling, obtaining an average accuracy of 0.86 on the test set, with performance on sub-tasks ranging from 0.68 to 0.98. These were the best results achieved at the competition overall. To allow lay people to easily use and see the value of machine learning for author profiling, we also built a web application on top our models.

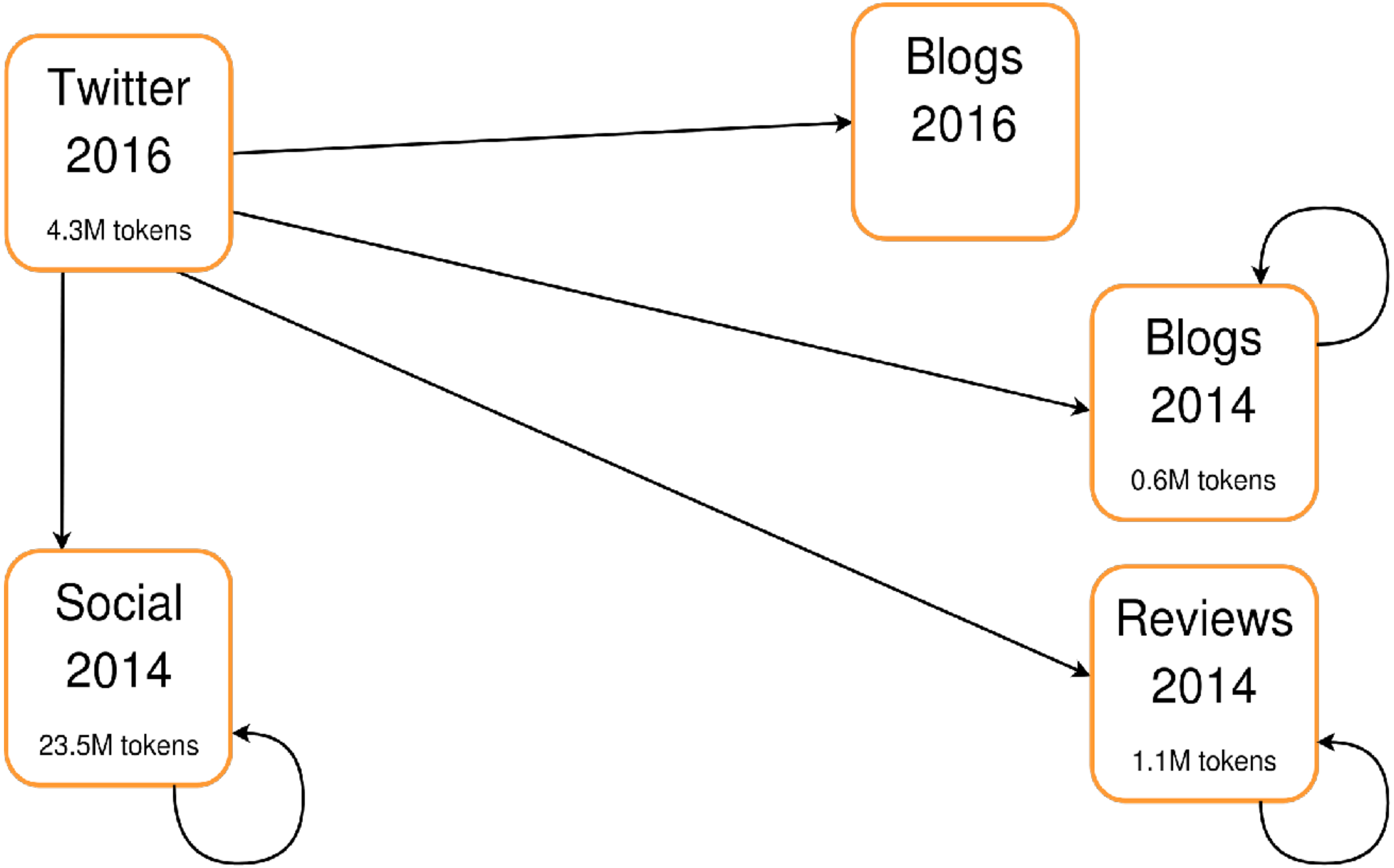
Keywords: author profiling · linear models · gender prediction · language variety identification · multitask learning

ARE WE REALLY DOING SO WELL?

PAN 2016 challenge on author profiling



Experimental Settings



Twitter 2016

4.3M tokens

Blogs 2016

Blogs 2014

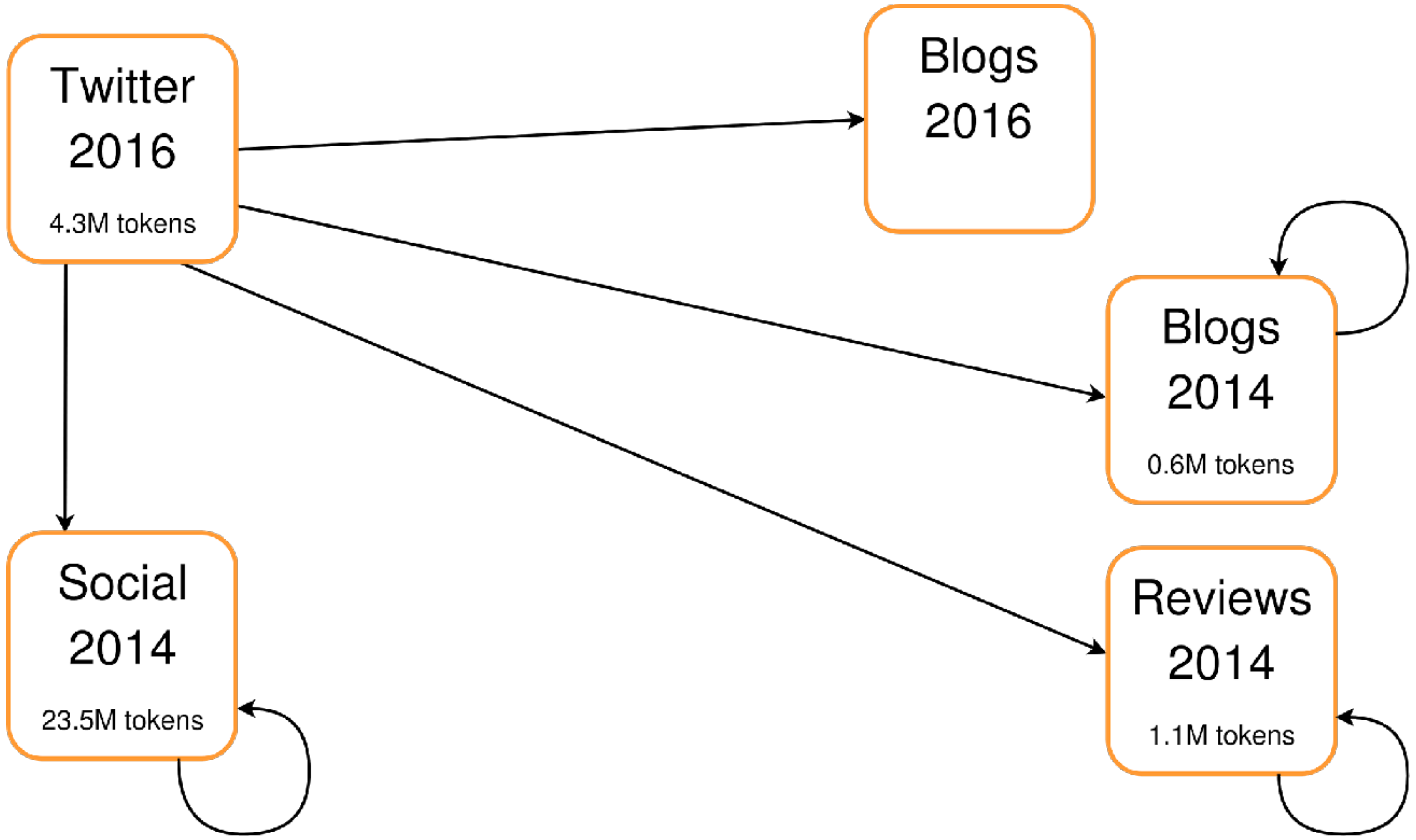
0.6M tokens

Social 2014

23.5M tokens

Reviews 2014

1.1M tokens



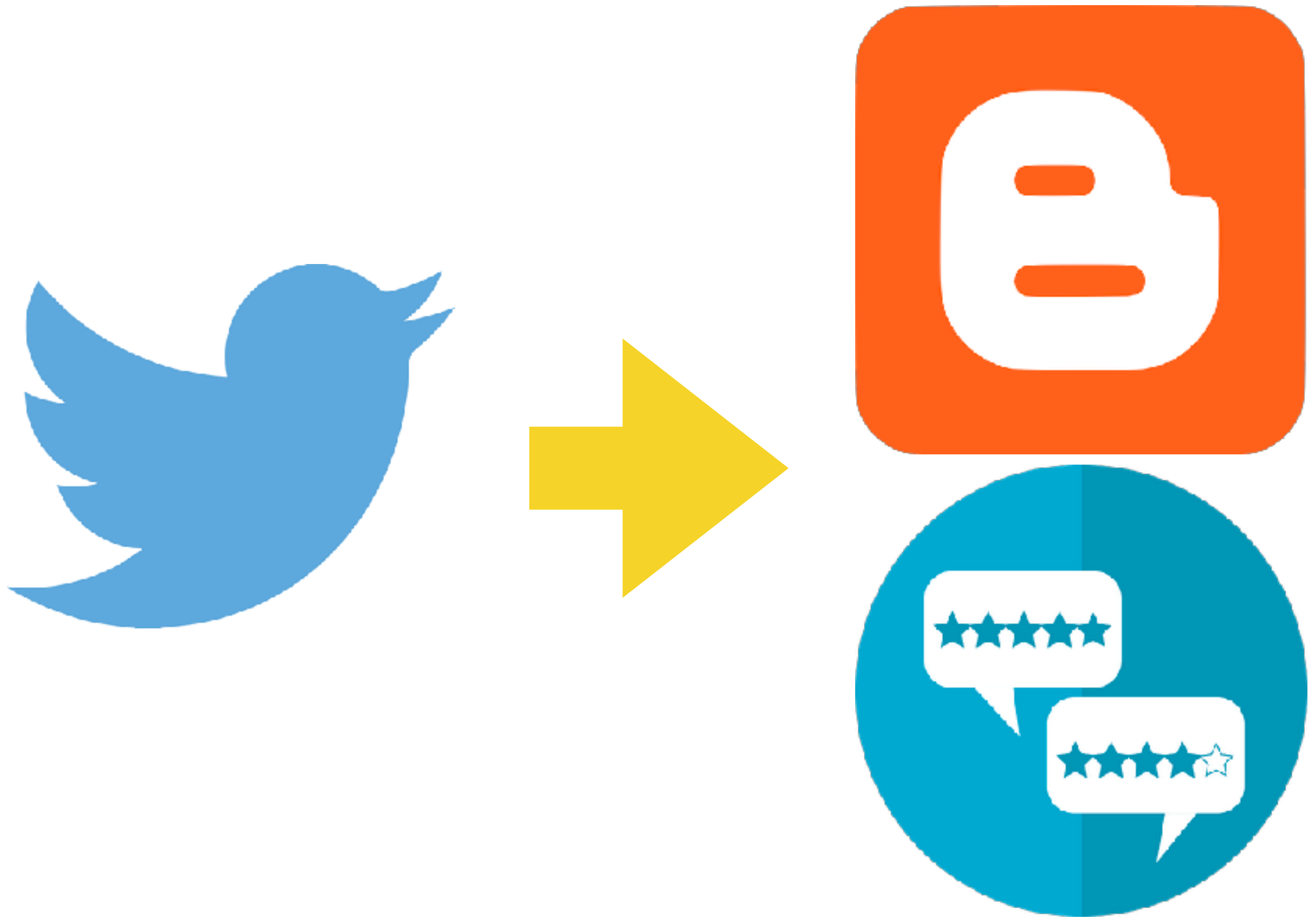
Twitter 2016
4.3M tokens

Blogs 2016

Blogs 2014
0.6M tokens

Reviews 2014
1.1M tokens

Social 2014
23.5M tokens



Training	Test	Average Acc.
Twitter 2016	Blogs 2016 (test)	0.6157
Twitter 2016	Blogs 2014	0.5936
Twitter 2016	Reviews 2014	0.3689
Twitter 2016	Social Media 2014	0.4240
Blogs 2014	Cross-Validation	0.5409
Reviews 2014	Cross-Validation	0.4881
Social Media 2014	Cross-Validation	0.4507

Cross-genre Insights

(size matters, quality matters, but in any case:)
Cross-genre profiling is hard

Reducing the help of lexical information yields lower scores

Cross-genre Insights

Cross-genre profiling is hard

Reducing the help of lexical information yields lower scores

IS THIS SUCH A BAD THING?

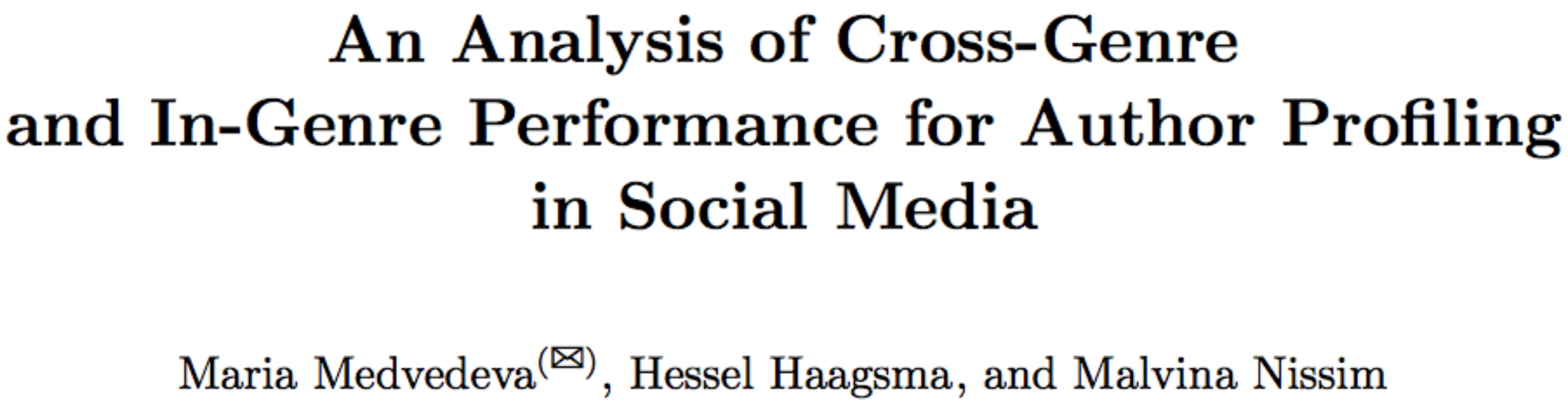
Cross-genre Insights

Cross-genre profiling is hard

Reducing the help of lexical information yields lower scores

IS THIS SUCH A BAD THING?

NO!



An Analysis of Cross-Genre and In-Genre Performance for Author Profiling in Social Media

Maria Medvedeva(✉), Hessel Haagsma, and Malvina Nissim

Proceedings of CLEF 2017

“A more systematic cross-genre evaluation setting is needed, controlling for various confounding factors: size, time, data quality”

GxG
Home
Submission
Task
Data
Dates
Related
Organisers

GxG

Cross-Genre Gender Prediction in Italian Shared Task at EVALITA 2018

- Twitter
- YouTube
- Journalism
- Personal Diaries
- Children's Writings

Training Data available: https://github.com/malvinanissim/gxg/tree/master/Data/Training

Overview of Datasets: https://github.com/malvinanissim/gxg/wiki/Data-description

GxG (Gender X-Genre) is a task on author profiling (in terms of gender) on Italian texts, with a specific focus on cross-genre performance, organised as part of EVALITA 2018

Cross-genre Insights

Cross-genre profiling is hard

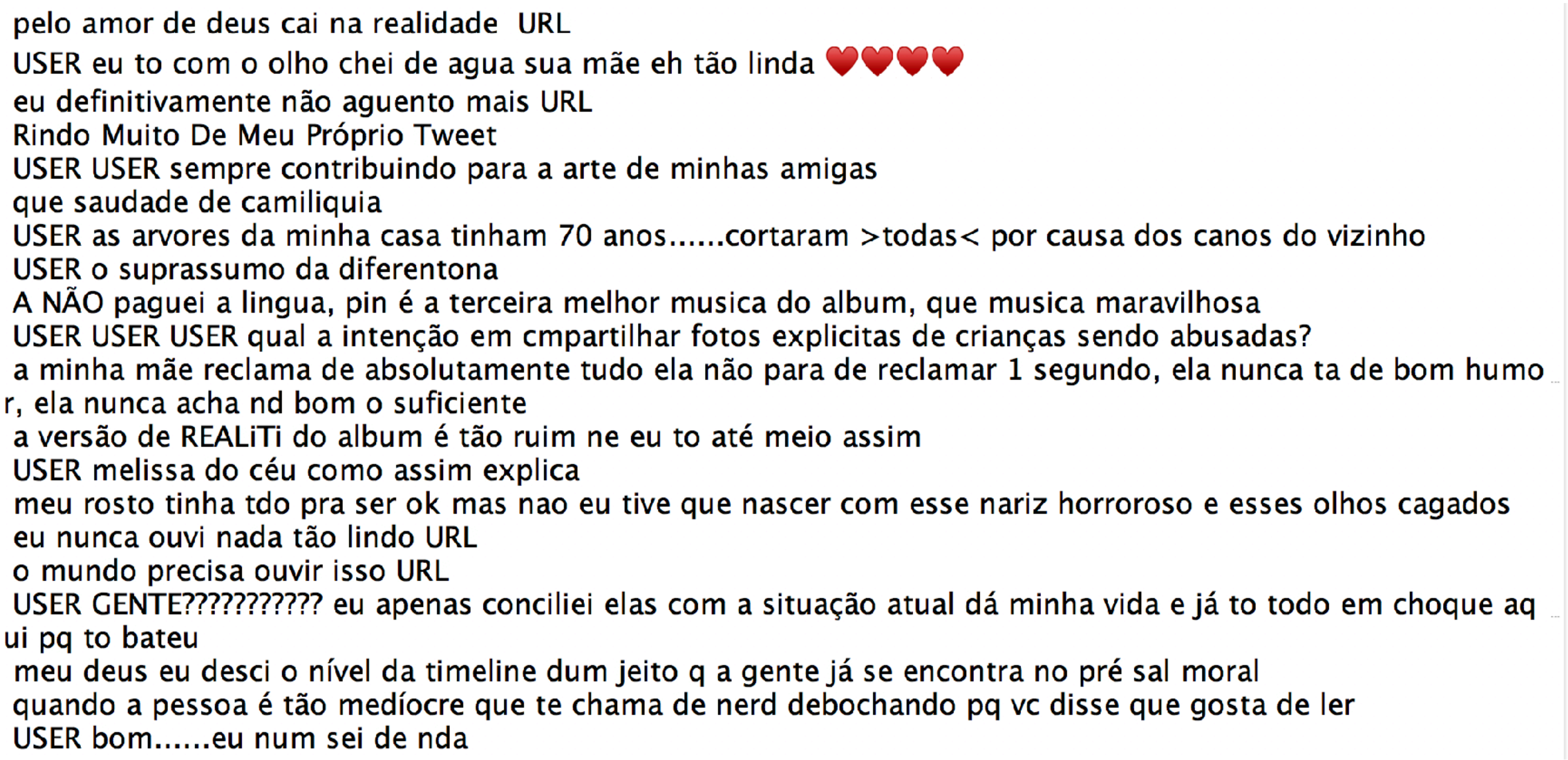
Reducing the help of lexical information yields lower scores

IS THIS SUCH A BAD THING?

NO!

being “forced” to abstract away from the lexicon might yield interesting insights (though lower scores)

male or female?



pelo amor de deus cai na realidade URL

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 nd bom o suficiente

a versão de REALiTi do album é tão ruim ne eu to até meio assim

USER melissa do céu como assim explica

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

male or female?

USER estavas cmg quase todos os dias babe
Vou mas é estudar e colar no spotify q é o q faço melhor
Pq as minhas séries favoritas são de sci fi, adoro hp, hunger games, divergente, avengers, avatar etc.
Ai n posso, vou te dar um n para n passares com três sims
De que estás à espera? Pulseiras+info cmg! 👼👼👼 URL
Digam me a cola zero tem cafeína na mesma é q secalhar é por andar a beber coca cola q n durmo pq eu n costumo beber,só agua ou iced tea
Siga acabar o trabalho mais estúpido e sem sentido q já tive, sim pq a minha explicadora passa se só de o ver
Mas a gopro pelo menos eu vou ter nem quero saber, nem q tenha q assaltar um banco para ter o resto do dinheiro
Quer dizer é mais ela a aturar me a mim but whatever,
Eu não seguia a ana Isabel (?)
Wtf isto é um programa de canto ou um desfile?!
Quando descobres q vais ter q todos os meses durante três anos apanhar uma injecao 😱😨 quando virem o meu braço ainda pensam q me drogo 😂😂😂
É q tinha bues a falar sobre álcool, drogas, jogos online e a dinheiro e eu tava sempre a responder q nunca tinha feito essas cenas
Depois o pessoal junta se todo e oferece me o capacete da vespa no aniversário ahahah
Agora viciei em pll, nunca tive curiosidade de ver e tp no dia em q tava tudo à espera para saber quem é o A eu a comecar ver 😂😂😂
Primeiro era fazes anos em fevereiro m só te deixo começar em maio e agora nem isso deixa tou capaz de a esganar!
HELP tou a ficar doente 😱😨 tou com altas dores de garganta
Ui uma de GoT aparece em HP 😱
USER ESQUECE ELA DISSE QUE ELE TAVA COM VARICELA AHAHAHAHAHAH
Tá aqui uma pessoa a estudar e ataca me até fiquei a largar sangue crg

male or female?

você é linda sim, onda do mar do amor que bateu em mim
e nós somos seus amiguinhosss os backyardigans
será que é bom ir no shopping com o cabelo pronto? nem seii
USER não aguento +
eu faço muita careta andando na rua, que vergonha
to rindo que nem idiota
USER USER esse cabelo amo, vale milhões
e nem é engraçado
vontade de entrar na conversa e falar ""ninguém liga""
a decoração de natal na disney tá lindaa
melhor turma que já tive
tem varias fotos zoadas dos outros no meu celular 🙄
de futebol ainda
mas se ele te merecesse não estaria aquii
minha mãe já quer ir embora
USER chegando na festa da luana URL
me segue de segunda a sexta feira
mó frio com esse ar ligado
tem duas coisas pra fazer hoje mas nenhuma é tão legal assim
gosto mais de geral do que brasil kaka

TESCO
fresh
bleach
24
KILLS GERMS
CUTS THROUGH DIRT & GRIME
THICK CLINGING FORMULATION

Bleaching text

`a bag of Doritos for lunch!` 😎

Bleaching text

a bag of Doritos for lunch! 😎

↓

01 03 02 07 03 06 01

4 2 4 0 4 1 0

v cvc vc cvcvcvc cvc cvccco o

L LL LL ULL LL LLX X

W W W W W WP J

W W W W W W! 😎

Bleaching text

a bag of Doritos for lunch! 😎

↓

Length (char): 01 03 02 07 03 06 01

Bleaching text

a bag of Doritos for lunch! 😎

Length (char): 01 03 02 07 03 06 01

Binned frequency: 4 2 4 0 4 1 0

Bleaching text

a bag of Doritos for lunch! 😎

↓

Length (char): 01 03 02 07 03 06 01

Binned frequency: 4 2 4 0 4 1 0

Vowel/Consonant: v cvc vc cvcvcvc cvc cvccco o

Bleaching text

a bag of Doritos for lunch! 😎

```
Length (char): 01 03 02 07 03 06 01

Binned frequency: 4  2  4  0  4  1  0

Vowel/Consonant: v cvc vc cvcvcvc cvc cvccco o

Lower/Upper/Other: L LL LL ULL LL LLX X
```

Bleaching text

a bag of Doritos for lunch! 😎

⬇

Length (char): 01 03 02 07 03 06 01

Binned frequency: 4 2 4 0 4 1 0

Vowel/Consonant: v cvc vc cvcvcvc cvc cvccco o

Lower/Upper/Other: L LL LL ULL LL LLX X

Punctuation (Aggr): W W W W W WP J

Bleaching text

a bag of Doritos for lunch! 😎

Length (char): 01 03 02 07 03 06 01

Binned frequency: 4 2 4 0 4 1 0

Vowel/Consonant: v cvc vc cvcvcvc cvc cvccco o

Lower/Upper/Other: L LL LL ULL LL LLX X

Punctuation (Aggr): W W W W W WP J

Punctuation (Cons): W W W W W W! 😎

Bleaching text

a bag of Doritos for lunch! 😎

↓

Length (char): 01 03 02 07 03 06 01

Binned frequency: 4 2 4 0 4 1 0

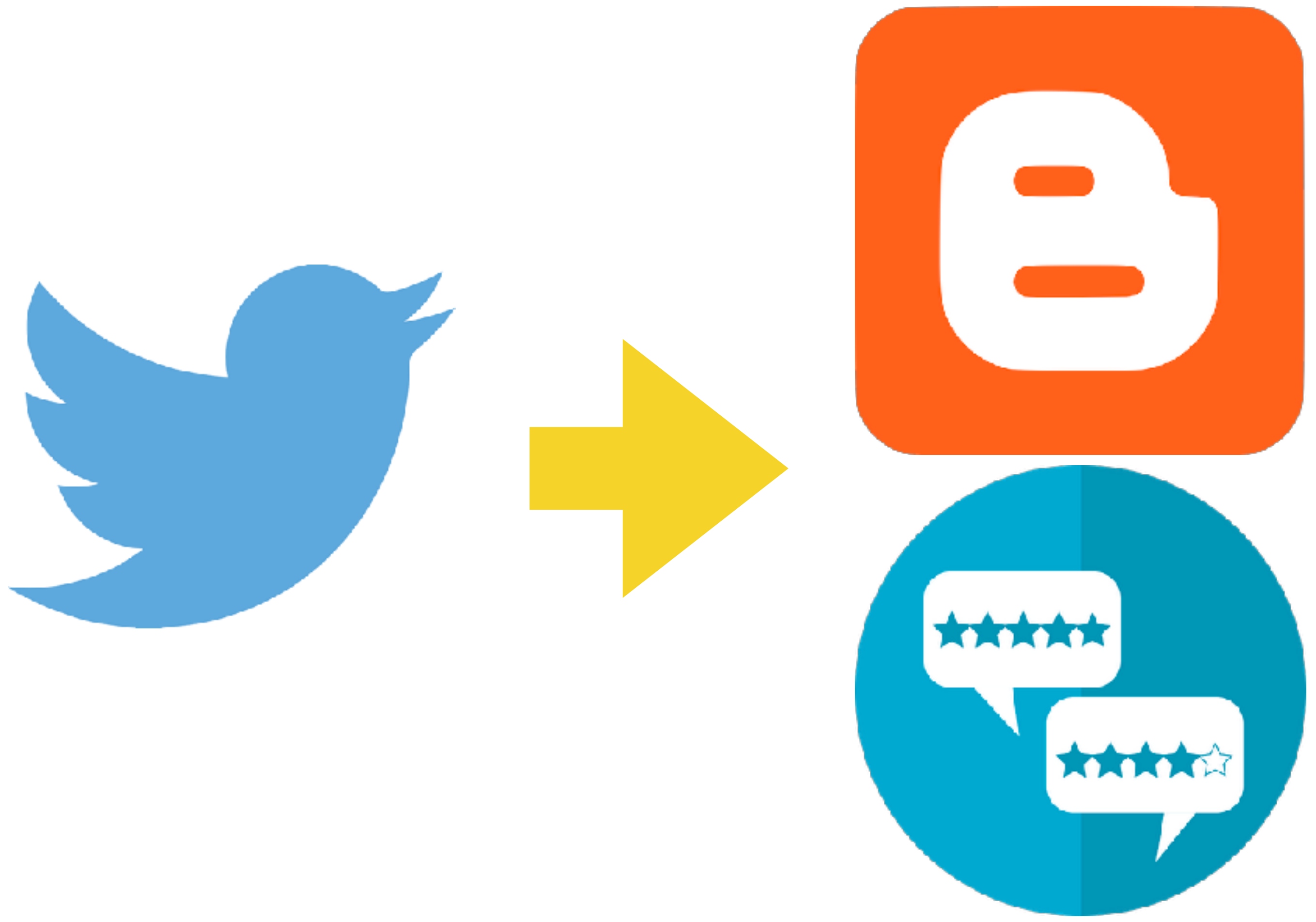
Vowel/Consonant: v cvc vc cvcvcvc cvc cvccco o

Lower/Upper/Other: L LL LL ULL LL LLX X

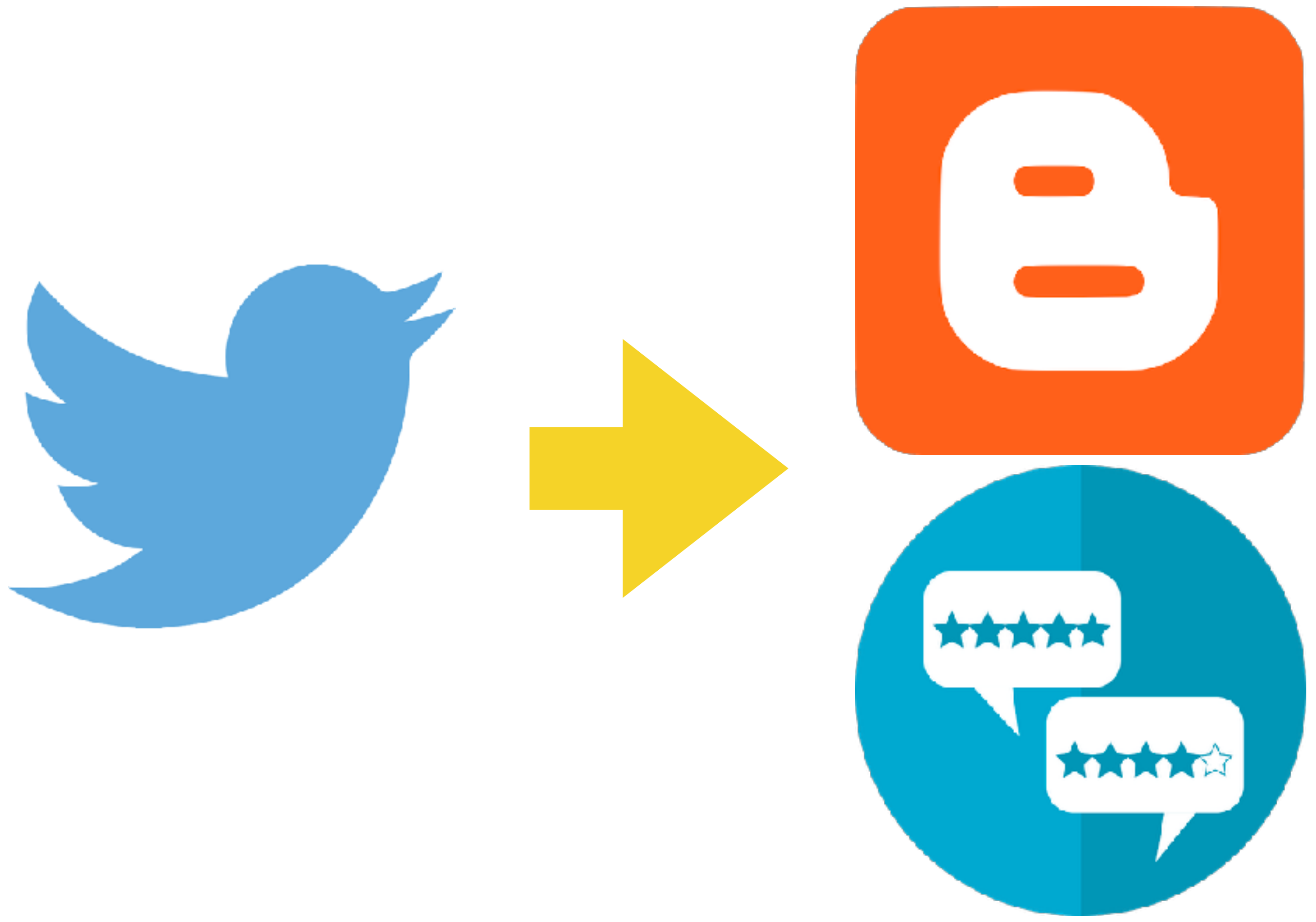
Punctuation (Aggr): W W W W W WP J

Punctuation (Cons): W W W W W W! 😎

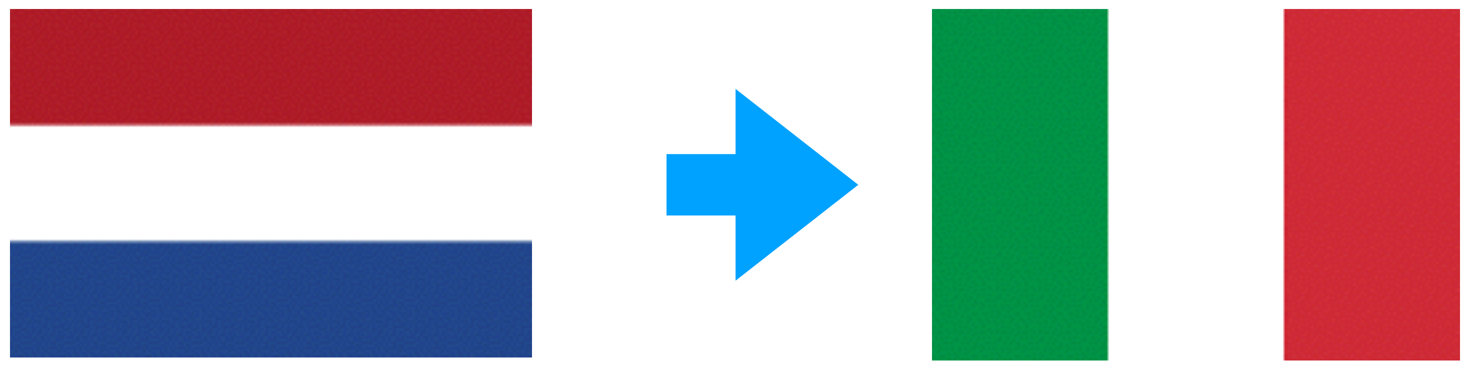
Cross-genre gender prediction



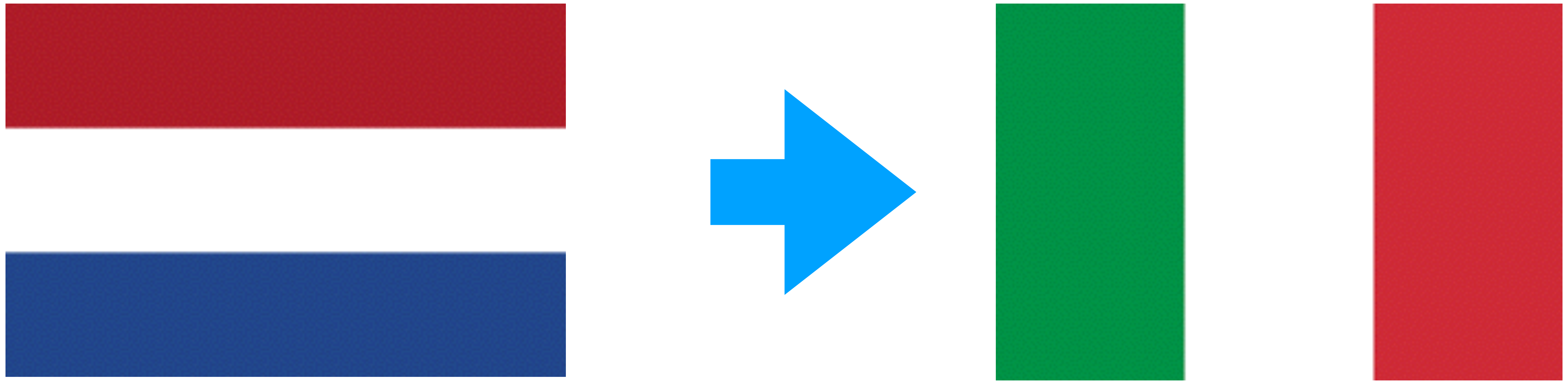
Cross-genre gender prediction



Cross-language gender prediction



Cross-language gender prediction



Cross-language Approaches

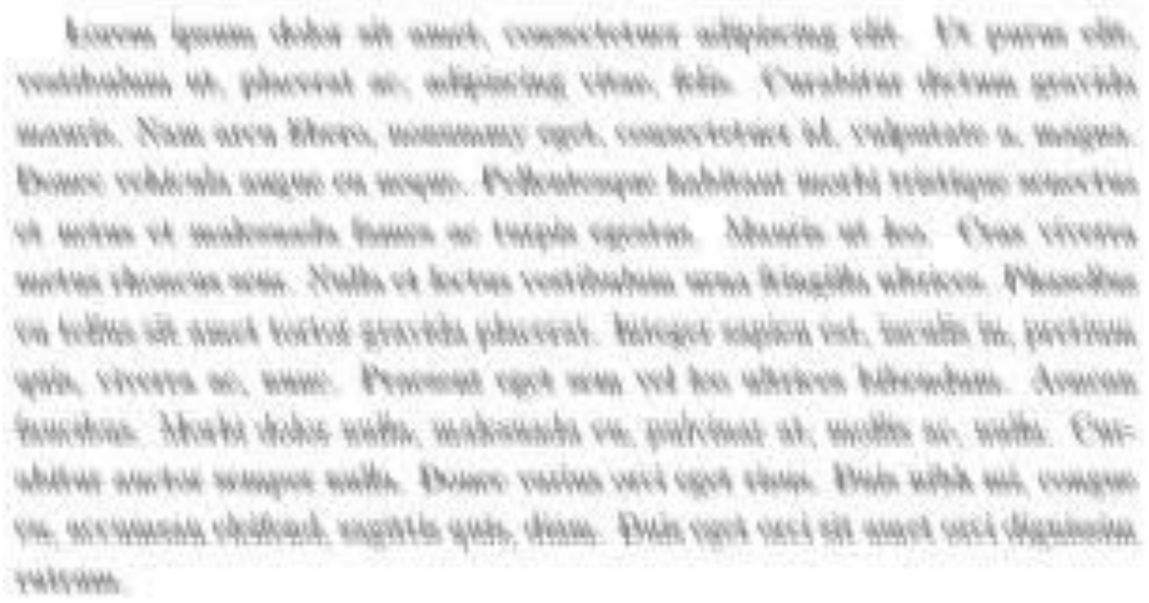
(Ljubešic et al. , 2017)

?..
@
#
http

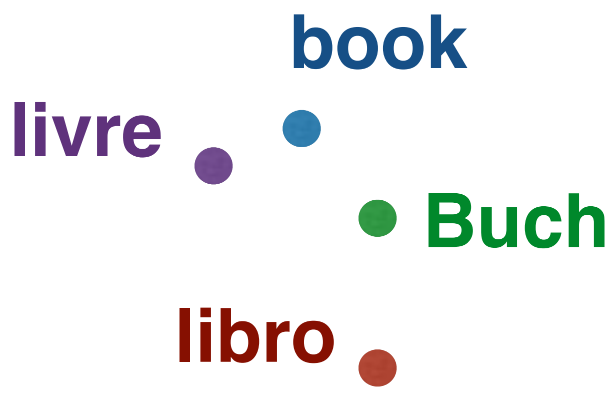
non-linguistic/
meta
features

open-vocabulary <> stylometry

abstract features



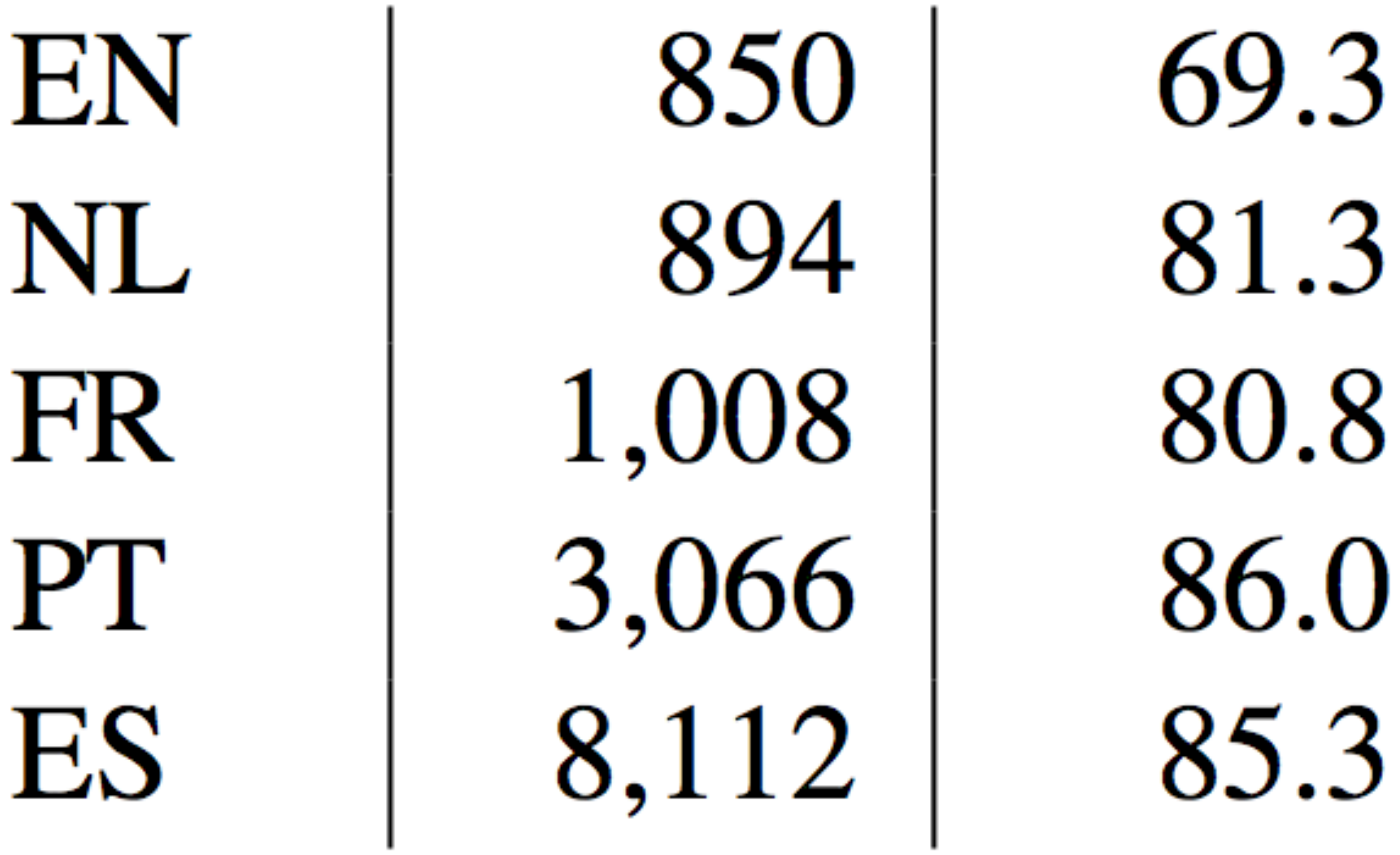
multilingual word
embeddings



Experimental Setup

- TwiSty corpus (ES, FR, NL, PT) + English, balanced for gender
- No preprocessing besides anonymizing URL and USER (not even tokenization)
- Lexical model: linear SVM with word and chr n-grams (winning system PAN 2017) on original tweets
- Bleached model: linear SVM with 5-grams (tuned via x-validation) on bleached tweets with concatenated representation
- In-language and Cross-language settings
  - in: 10fold x-validation
  - cross: one *other* language per time (average) and all *other* concatenated languages

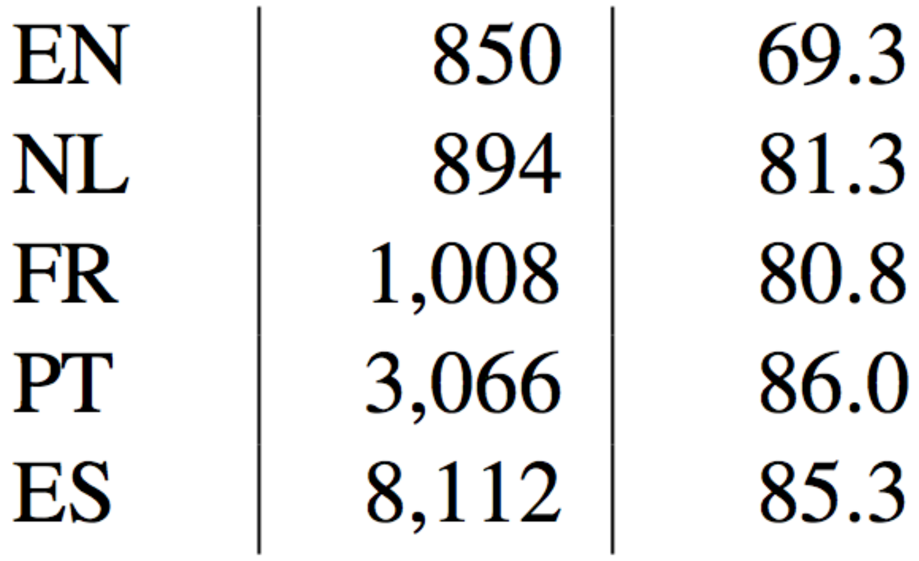
Lexical Model: In-language



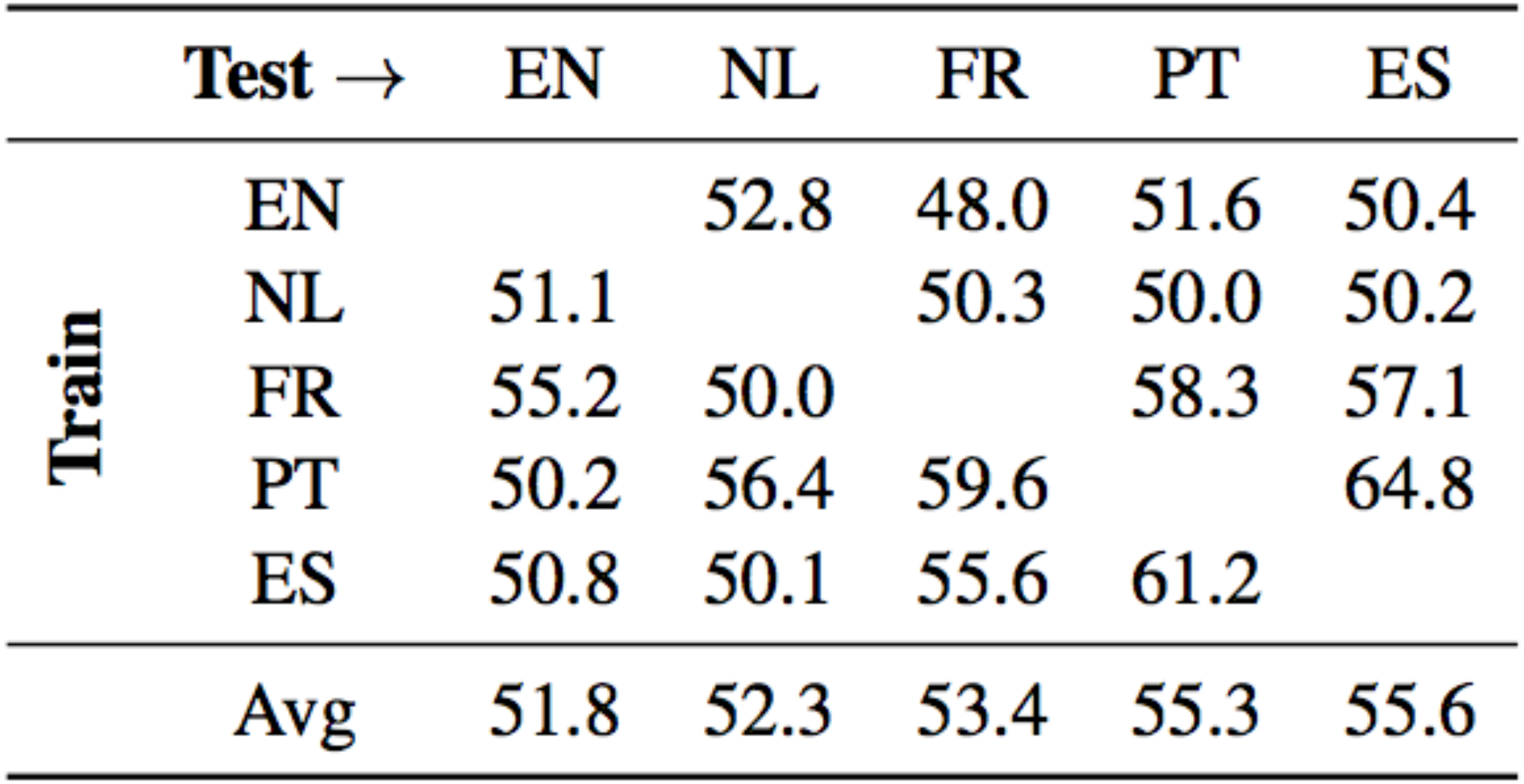
Lang	Users	Accuracy
EN	850	69.3
NL	894	81.3
FR	1,008	80.8
PT	3,066	86.0
ES	8,112	85.3

(one user = 200 tweets)

Lexical Model: Cross-language

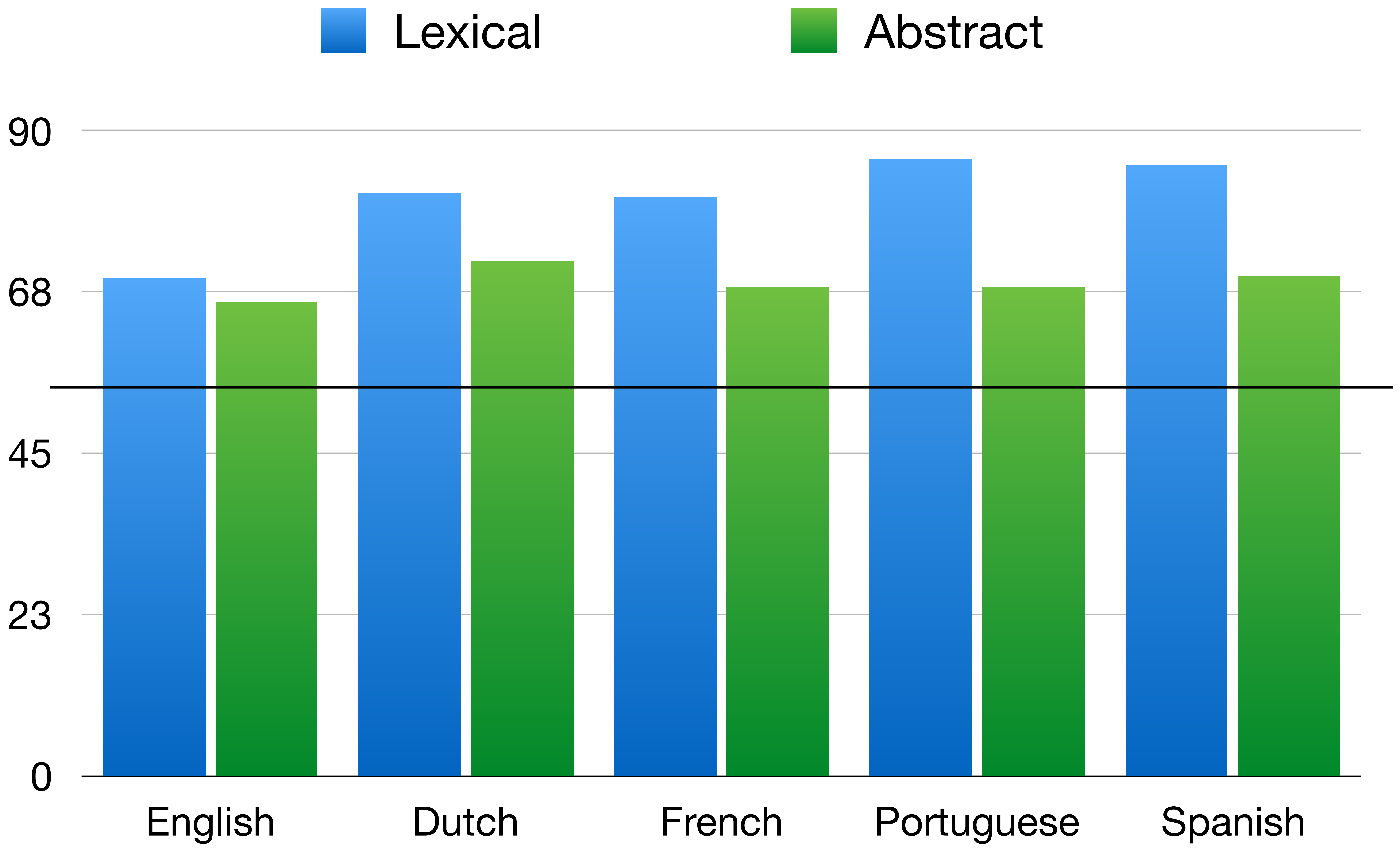


EN	850	69.3
NL	894	81.3
FR	1,008	80.8
PT	3,066	86.0
ES	8,112	85.3

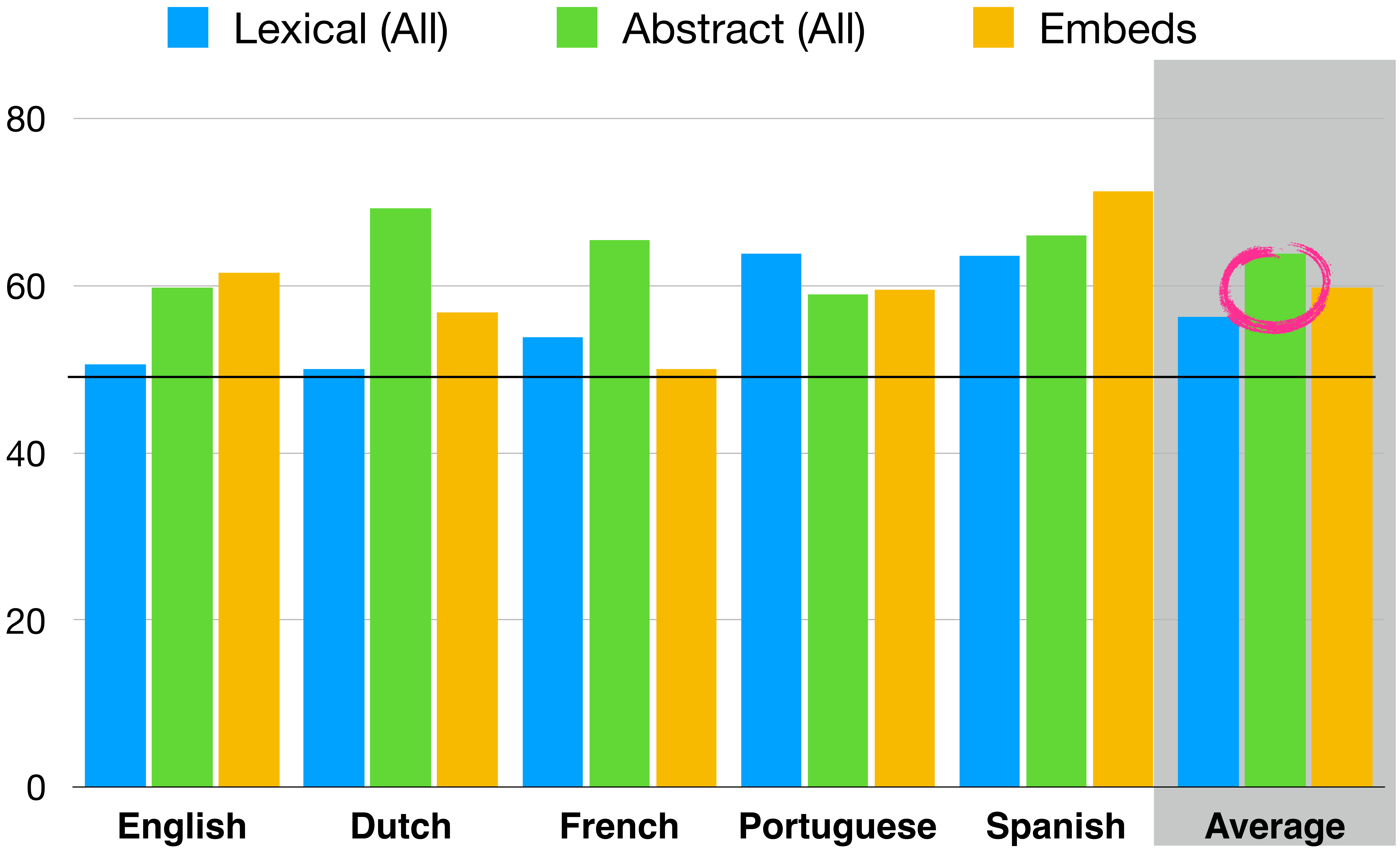


	Test →	EN	NL	FR	PT	ES
Train	EN		52.8	48.0	51.6	50.4
	NL	51.1		50.3	50.0	50.2
	FR	55.2	50.0		58.3	57.1
	PT	50.2	56.4	59.6		64.8
	ES	50.8	50.1	55.6	61.2	
	Avg	51.8	52.3	53.4	55.3	55.6

Bleached Model: In-language

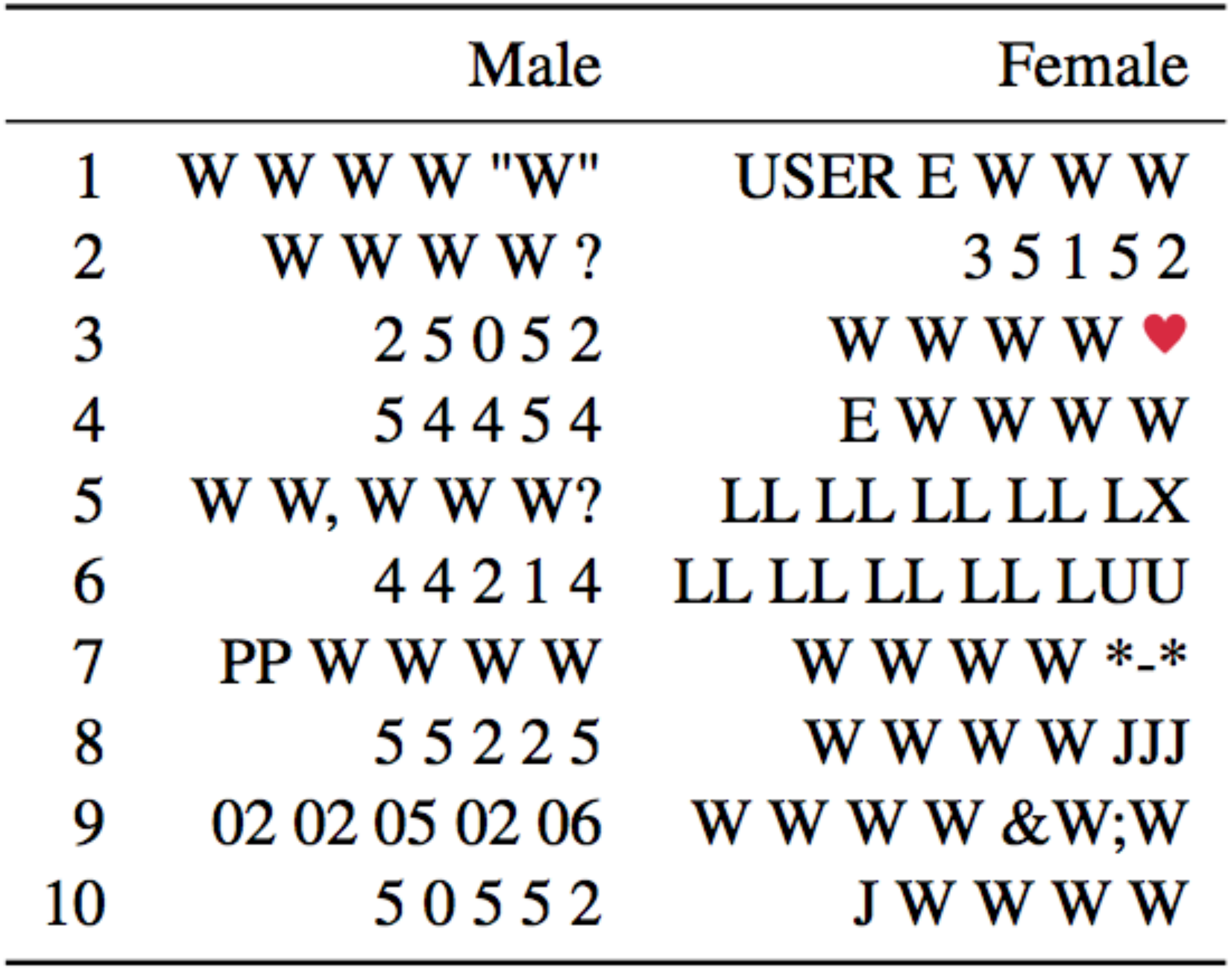


Bleached Model: Cross-language



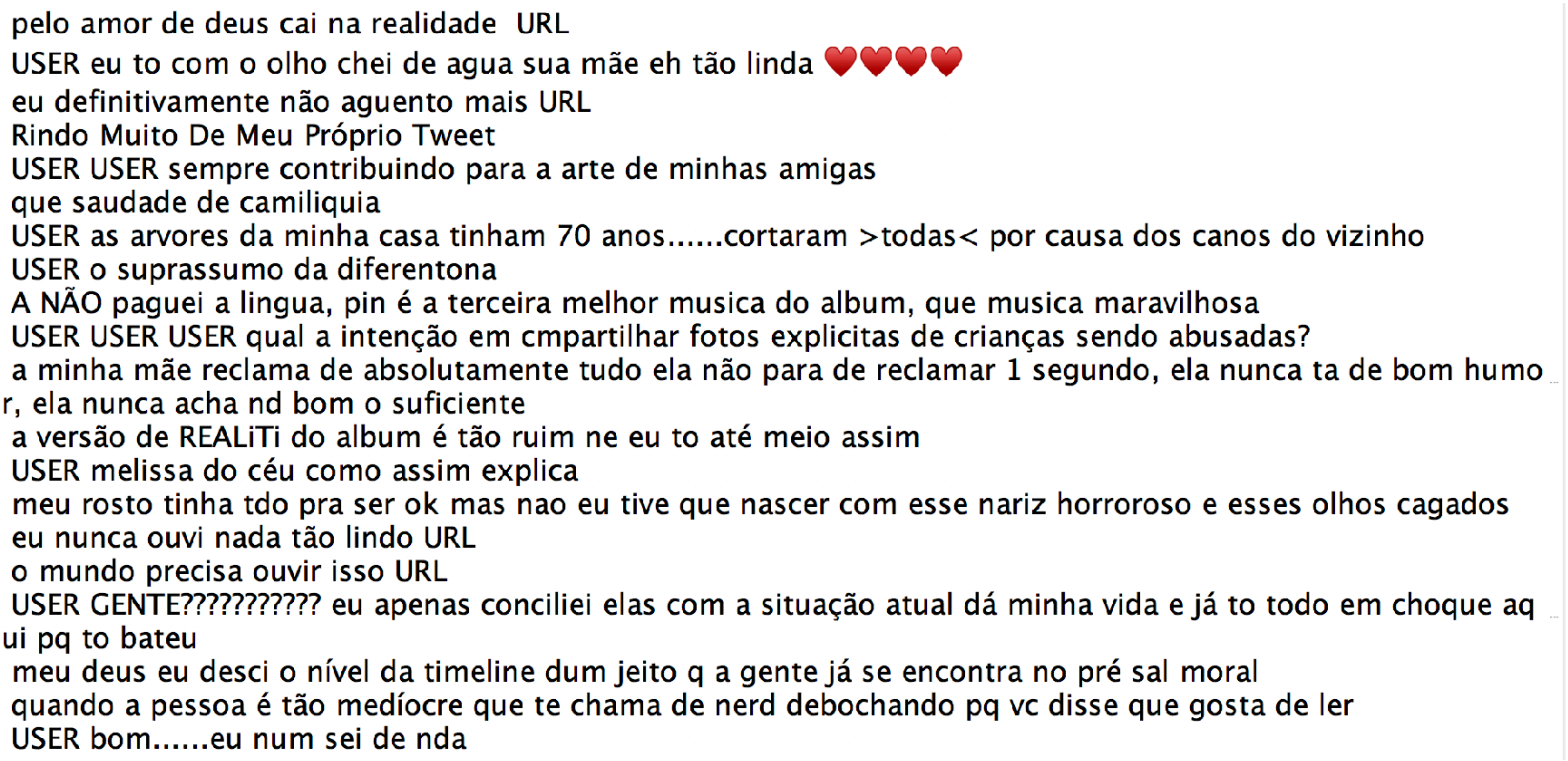
NB: huge in-domain specific embeddings necessary!

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

male or female?



pelo amor de deus cai na realidade URL
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 nd bom o suficiente
a versão de REALiTi do album é tão ruim ne eu to até meio assim
USER melissa do céu como assim explica
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

male or female?

USER estavas cmg quase todos os dias babe
Vou mas é estudar e colar no spotify q é o q faço melhor
Pq as minhas séries favoritas são de sci fi, adoro hp, hunger games, divergente, avengers, avatar etc.
Ai n posso, vou te dar um n para n passares com três sims
De que estás à espera? Pulseiras+info cmg! 👼👼👼 URL
Digam me a cola zero tem cafeína na mesma é q secalhar é por andar a beber coca cola q n durmo pq eu n costumo beber,só agua ou iced tea
Siga acabar o trabalho mais estúpido e sem sentido q já tive, sim pq a minha explicadora passa se só de o ver
Mas a gopro pelo menos eu vou ter nem quero saber, nem q tenha q assaltar um banco para ter o resto do dinheiro
Quer dizer é mais ela a aturar me a mim but whatever,
Eu não seguia a ana Isabel (?)
Wtf isto é um programa de canto ou um desfile?!
Quando descobres q vais ter q todos os meses durante três anos apanhar uma injecao 😱😨 quando virem o meu braço ainda pensam q me drogo 😂😂😂
É q tinha bues a falar sobre álcool, drogas, jogos online e a dinheiro e eu tava sempre a responder q nunca tinha feito essas cenas
Depois o pessoal junta se todo e oferece me o capacete da vespa no aniversário ahahah
Agora viciei em pll, nunca tive curiosidade de ver e tp no dia em q tava tudo à espera para saber quem é o A eu a comecar ver 😂😂😂
Primeiro era fazes anos em fevereiro m só te deixo começar em maio e agora nem isso deixa tou capaz de a esganar!
HELP tou a ficar doente 😱😨 tou com altas dores de garganta
Ui uma de GoT aparece em HP 😱
USER ESQUECE ELA DISSE QUE ELE TAVA COM VARICELA AHAHAHAHAHAH
Tá aqui uma pessoa a estudar e ataca me até fiquei a largar sangue crg

male or female?

você é linda sim, onda do mar do amor que bateu em mim
e nós somos seus amiguinhosss os backyardigans
será que é bom ir no shopping com o cabelo pronto? nem seii
USER não aguento +
eu faço muita careta andando na rua, que vergonha
to rindo que nem idiota
USER USER esse cabelo amo, vale milhões
e nem é engraçado
vontade de entrar na conversa e falar ""ninguém liga""
a decoração de natal na disney tá lindaa
melhor turma que já tive
tem varias fotos zoadas dos outros no meu celular 🙄
de futebol ainda
mas se ele te merecesse não estaria aquii
minha mãe já quer ir embora
USER chegando na festa da luana URL
me segue de segunda a sexta feira
mó frio com esse ar ligado
tem duas coisas pra fazer hoje mas nenhuma é tão legal assim
gosto mais de geral do que brasil kaka

Human Model: Settings

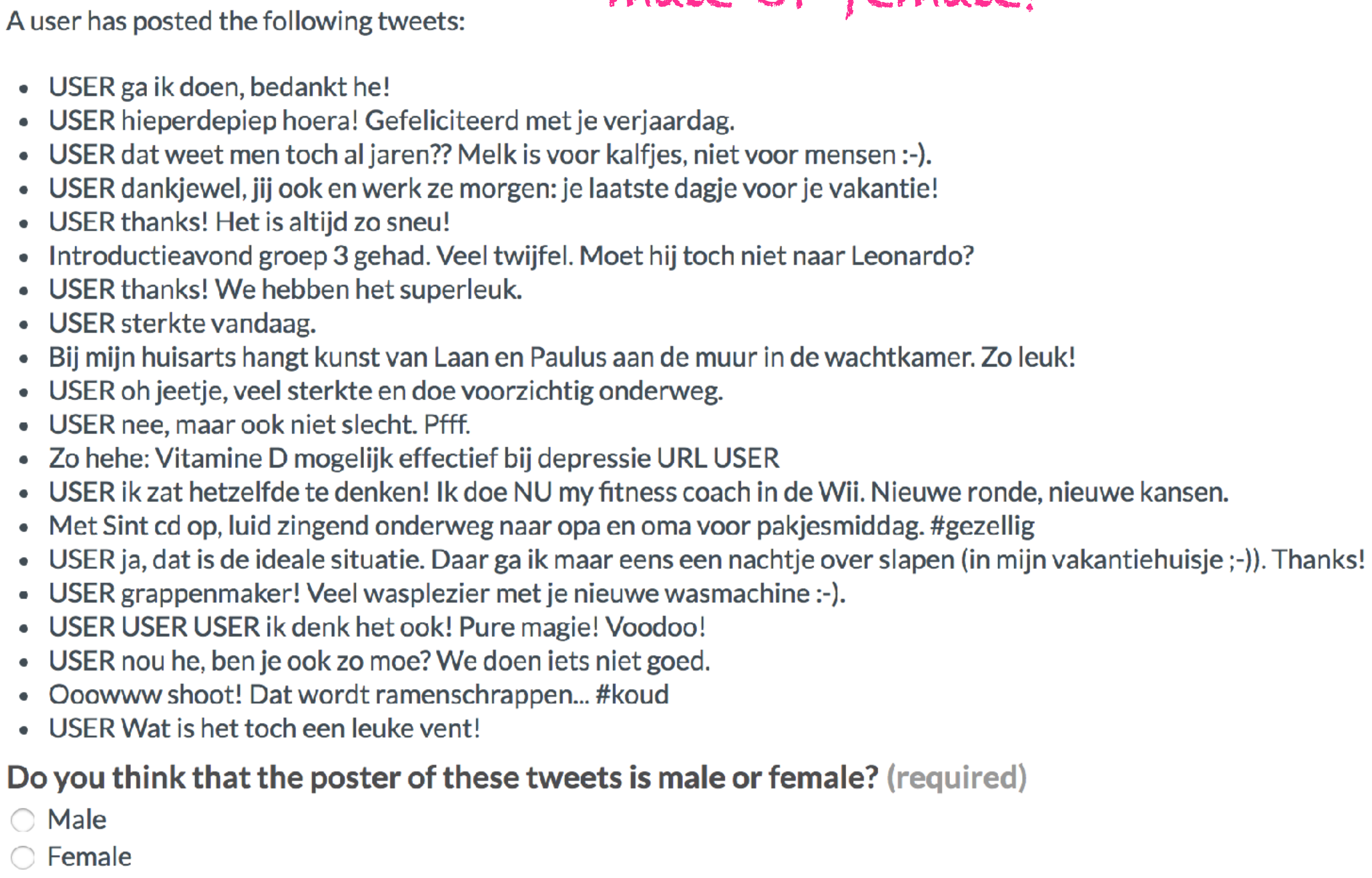
- Conditions:

  In-language: Dutch→Dutch

  Cross-language: Dutch→Portuguese
  French→Dutch

- 20 tweets, randomized answer key, 3 annotators

- Data balanced per gender, participants balanced by gender male or female?



A user has posted the following tweets:

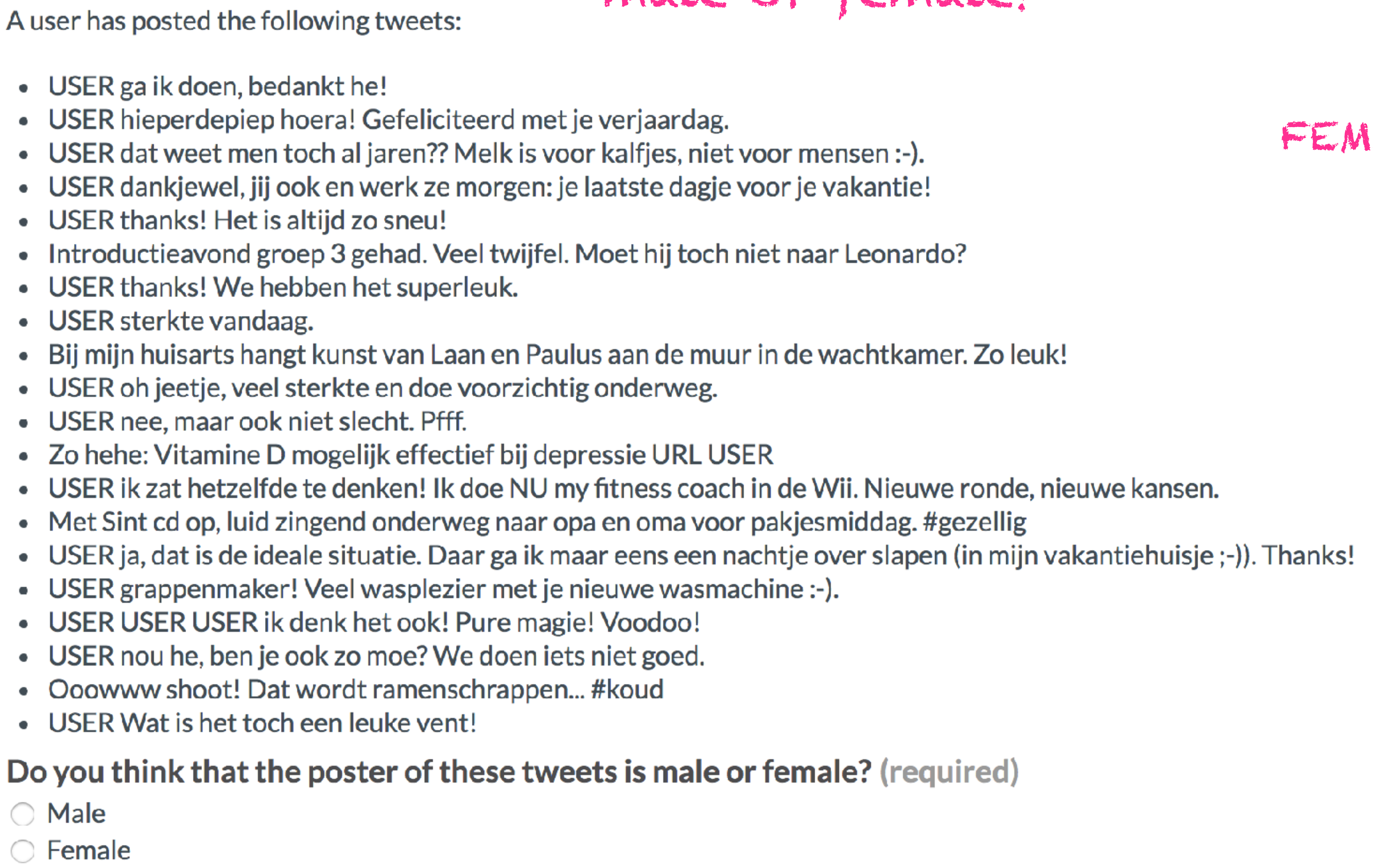
- USER ga ik doen, bedankt he!
- USER hieperdepiep hoera! Gefeliciteerd met je verjaardag.
- USER dat weet men toch al jaren?? Melk is voor kalfjes, niet voor mensen :-).
- USER dankjewel, jij ook en werk ze morgen: je laatste dagje voor je vakantie!
- USER thanks! Het is altijd zo sneu!
- Introductieavond groep 3 gehad. Veel twijfel. Moet hij toch niet naar Leonardo?
- USER thanks! We hebben het superleuk.
- USER sterkte vandaag.
- Bij mijn huisarts hangt kunst van Laan en Paulus aan de muur in de wachtkamer. Zo leuk!
- USER oh jeetje, veel sterkte en doe voorzichtig onderweg.
- USER nee, maar ook niet slecht. Pfff.
- Zo hehe: Vitamine D mogelijk effectief bij depressie URL USER
- USER ik zat hetzelfde te denken! Ik doe NU my fitness coach in de Wii. Nieuwe ronde, nieuwe kansen.
- Met Sint cd op, luid zingend onderweg naar opa en oma voor pakjesmiddag. #gezellig
- USER ja, dat is de ideale situatie. Daar ga ik maar eens een nachtje over slapen (in mijn vakantiehuisje ;-)). Thanks!
- USER grappenmaker! Veel wasplezier met je nieuwe wasmachine :-).
- USER USER USER ik denk het ook! Pure magie! Voodoo!
- USER nou he, ben je ook zo moe? We doen iets niet goed.
- Ooowww shoot! Dat wordt ramenschrappen... #koud
- USER Wat is het toch een leuke vent!

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

○ Male

○ Female

male or female?



A user has posted the following tweets:

- USER ga ik doen, bedankt he!
- USER hieperdepiep hoera! Gefeliciteerd met je verjaardag.
- USER dat weet men toch al jaren?? Melk is voor kalfjes, niet voor mensen :-).
- USER dankjewel, jij ook en werk ze morgen: je laatste dagje voor je vakantie!
- USER thanks! Het is altijd zo sneu!
- Introductieavond groep 3 gehad. Veel twijfel. Moet hij toch niet naar Leonardo?
- USER thanks! We hebben het superleuk.
- USER sterkte vandaag.
- Bij mijn huisarts hangt kunst van Laan en Paulus aan de muur in de wachtkamer. Zo leuk!
- USER oh jeetje, veel sterkte en doe voorzichtig onderweg.
- USER nee, maar ook niet slecht. Pfff.
- Zo hehe: Vitamine D mogelijk effectief bij depressie URL USER
- USER ik zat hetzelfde te denken! Ik doe NU my fitness coach in de Wii. Nieuwe ronde, nieuwe kansen.
- Met Sint cd op, luid zingend onderweg naar opa en oma voor pakjesmiddag. #gezellig
- USER ja, dat is de ideale situatie. Daar ga ik maar eens een nachtje over slapen (in mijn vakantiehuisje ;-)). Thanks!
- USER grappenmaker! Veel wasplezier met je nieuwe wasmachine :-).
- USER USER USER ik denk het ook! Pure magie! Voodoo!
- USER nou he, ben je ook zo moe? We doen iets niet goed.
- Ooowww shoot! Dat wordt ramenschrappen... #koud
- USER Wat is het toch een leuke vent!

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

○ Male

○ Female

FEMALE!

male or female?

A user has posted the following tweets:

- En deze hebben jullie ook nog tegoed: #StadshaardToday URL
- Was weer een erg leuk gesprek op VRT talkshow met Ivar ten Velde. Is Han Pape wellicht de Twentse Ischa Meijer?
- Honger! Hup hup naar mijn lunchafspraak USER Smaak
- Nederland toch niet toe aan een boy band... Enschede
- Vroeger gaf je het volk brood en spelen. Nu geef je ze 130km/u. URL #verkiezingen URL
- USER Sorry was de hele dag op pad. Lekker in de haven van Hengelo opnames maken: URL
- Hoe sneu ben je als je als volwassen man voetbalplaatjes staat te ruilen bij de grootgrutter? URL
- Aanstaande vrijdag te koop in de app store: de wandel- en fietsroutesb
- Kan geen toeval zijn; RT USER USER Het is perfect saunaweer #holterhof
- 'Mam, ik zag een man met bloemetjes aan zijn mandje, dat kan toch niet?' URL (gelukkig zag ze mijn crocs niet)
- Genieten van gitaar USER festival (@ Concordia Theater) URL
- Opbrengst van vandaag: URL + 2 gedouchte kidz #zontoday
- Zodadelijk eerst het snorren-overleg, dan naar Ruud van Palthehuis. Eens kijken of daar al wat Four Square points te vinden zijn...
- Zou God nu elke dag op 264 kanalen tegelijk naar een soort 'Tussen leven en dood' moeten kijken? Arme man.
- Volgens mij gaat het vooral over Jeroen deze keer #projectcatwalk
- USER Thanx! We nemen een dagje vrij vrijdag!
- Ik lees nu: "Feest vieren in de krant" URL
- USER wat is de stand?
- USER Precies! Al die onbelangrijkheid moet wel goed gedoseerd worden.
- Bord voor de kop van het CDA komt in Guinness WorldRecords 2010!

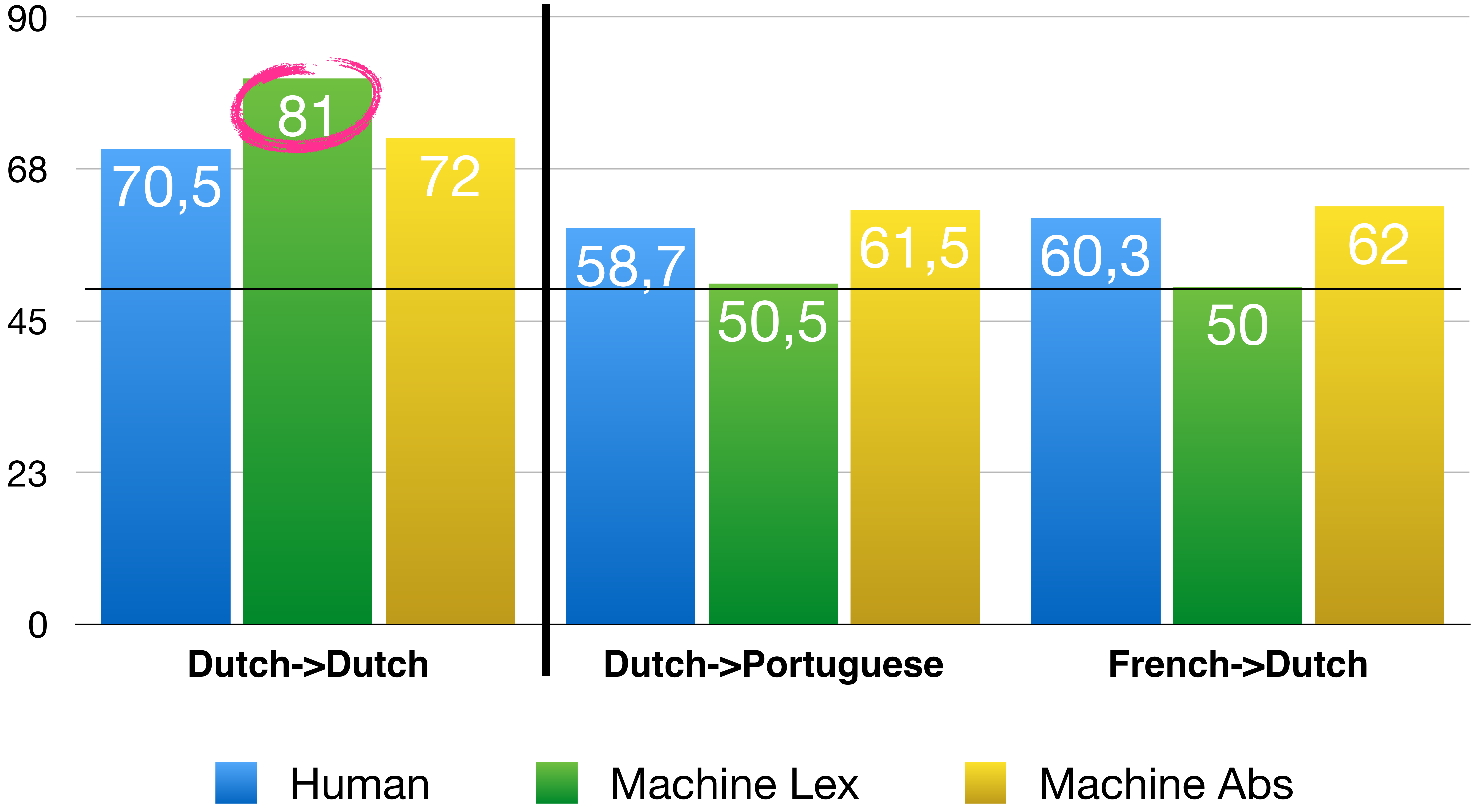
male or female?

A user has posted the following tweets:

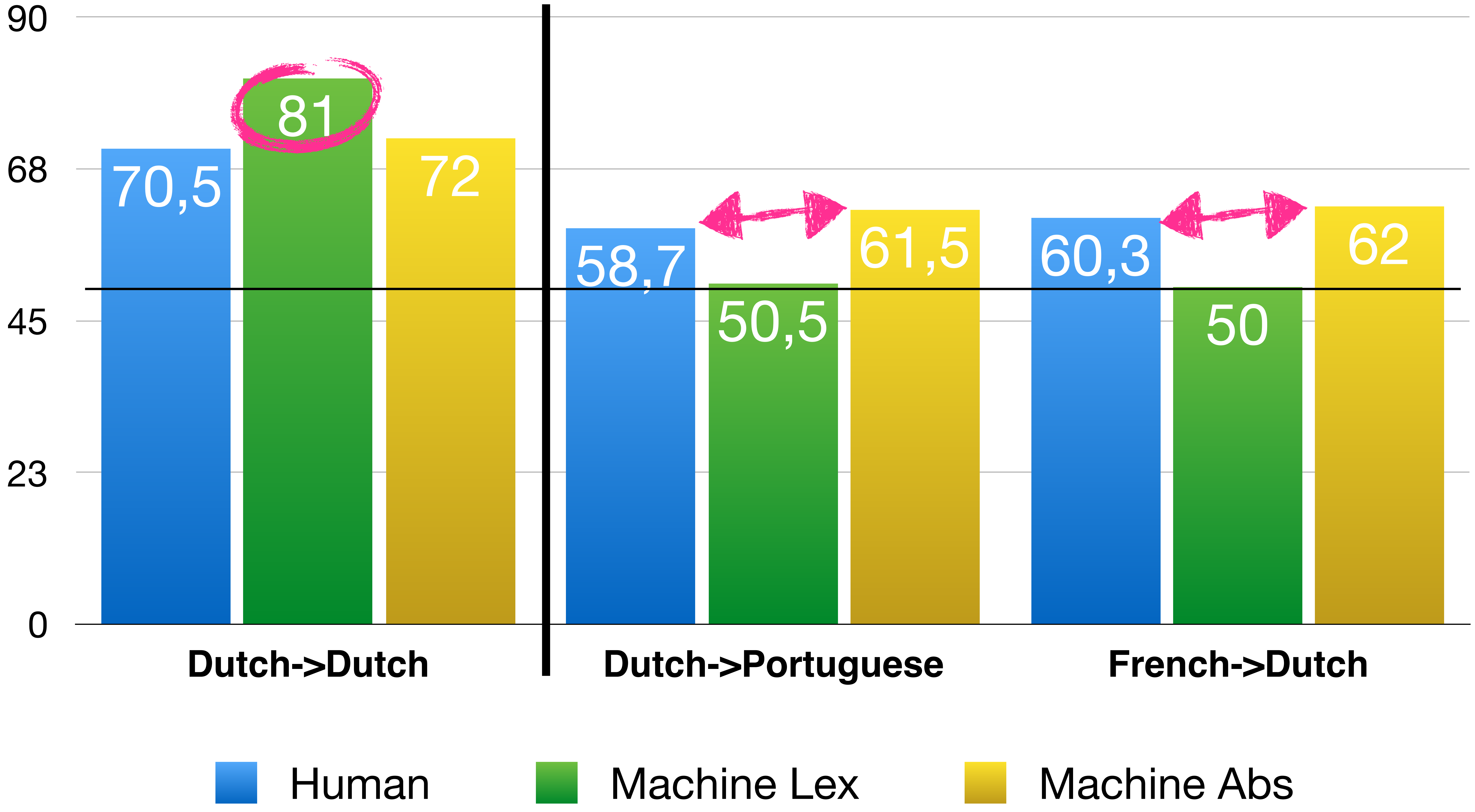
MALE!

- En deze hebben jullie ook nog tegoed: #StadshaardToday URL
- Was weer een erg leuk gesprek op VRT talkshow met Ivar ten Velde. Is Han Pape wellicht de Twentse Ischa Meijer?
- Honger! Hup hup naar mijn lunchafspraak USER Smaak
- Nederland toch niet toe aan een boy band... Enschede
- Vroeger gaf je het volk brood en spelen. Nu geef je ze 130km/u. URL #verkiezingen URL
- USER Sorry was de hele dag op pad. Lekker in de haven van Hengelo opnames maken: URL
- Hoe sneu ben je als je als volwassen man voetbalplaatjes staat te ruilen bij de grootgrutter? URL
- Aanstaande vrijdag te koop in de app store: de wandel- en fietsroutesb
- Kan geen toeval zijn; RT USER USER Het is perfect saunaweer #holterhof
- 'Mam, ik zag een man met bloemetjes aan zijn mandje, dat kan toch niet?' URL (gelukkig zag ze mijn crocs niet)
- Genieten van gitaar USER festival (@ Concordia Theater) URL
- Opbrengst van vandaag: URL + 2 gedouchte kidz #zontoday
- Zodadelijk eerst het snorren-overleg, dan naar Ruud van Palthehuis. Eens kijken of daar al wat Four Square points te vinden zijn...
- Zou God nu elke dag op 264 kanalen tegelijk naar een soort 'Tussen leven en dood' moeten kijken? Arme man.
- Volgens mij gaat het vooral over Jeroen deze keer #projectcatwalk
- USER Thanx! We nemen een dagje vrij vrijdag!
- Ik lees nu: "Feest vieren in de krant" URL
- USER wat is de stand?
- USER Precies! Al die onbelangrijkheid moet wel goed gedoseerd worden.
- Bord voor de kop van het CDA komt in Guinness WorldRecords 2010!

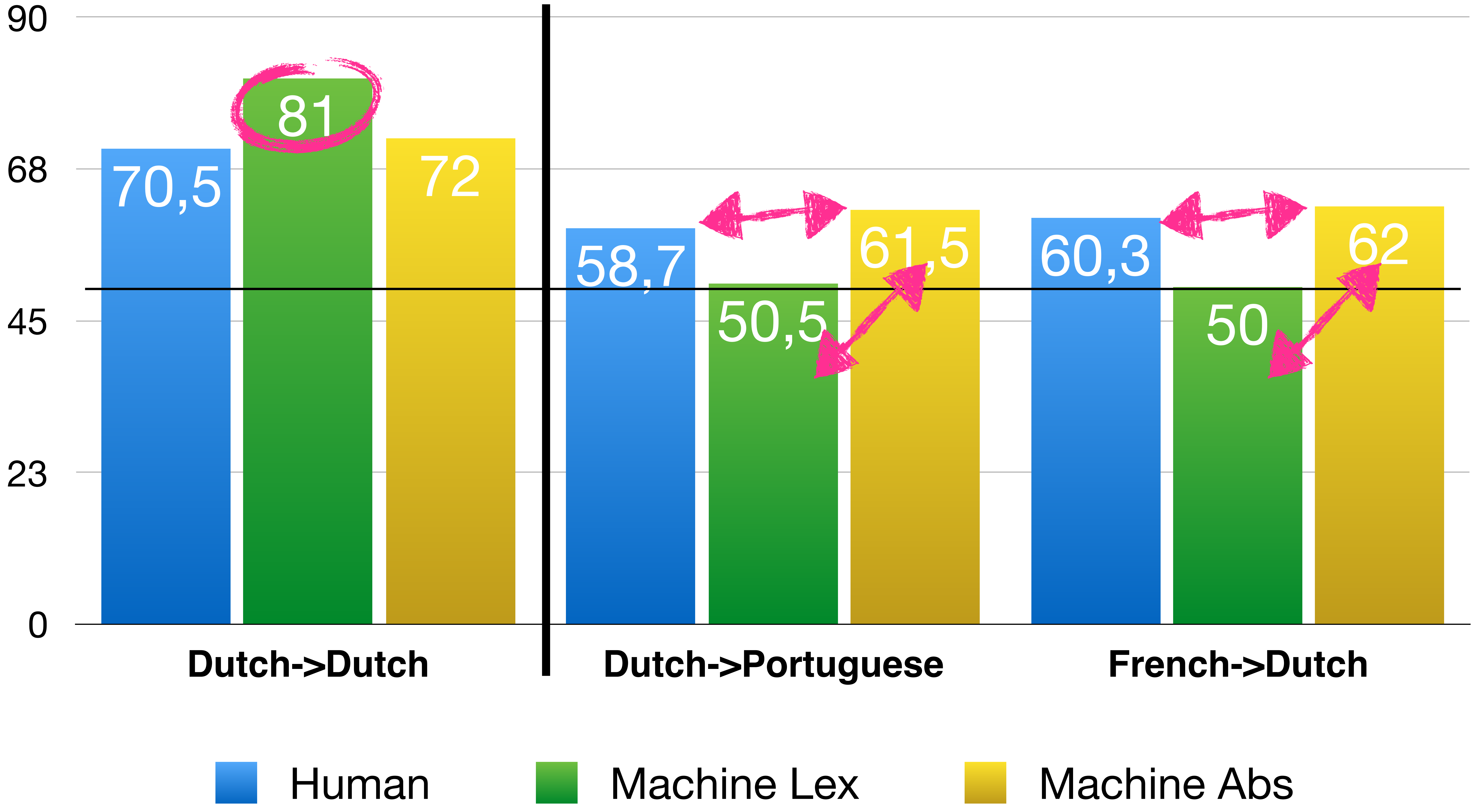
Comparative Results



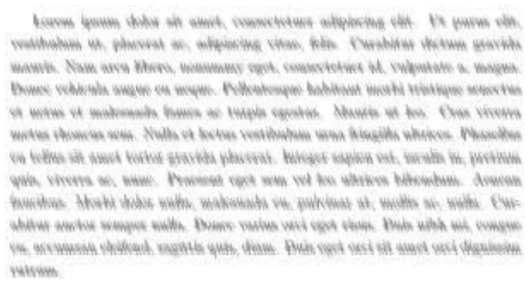
Comparative Results



Comparative Results

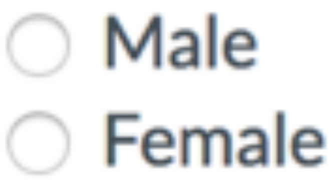


Main take-aways



- Bleaching text into abstract features is surprisingly effective, in-language (in-dataset) lexical features are best

- Abstract features transfer across 6 Indo-European languages, outperforming multilingual embeddings



○ Male
○ Female

- Humans can do cross-lingual gender prediction with 70.5% accuracy

What to do next

★Bleaching cross-genre

★Bleaching cross-datasets (use the PAN 2017 data)

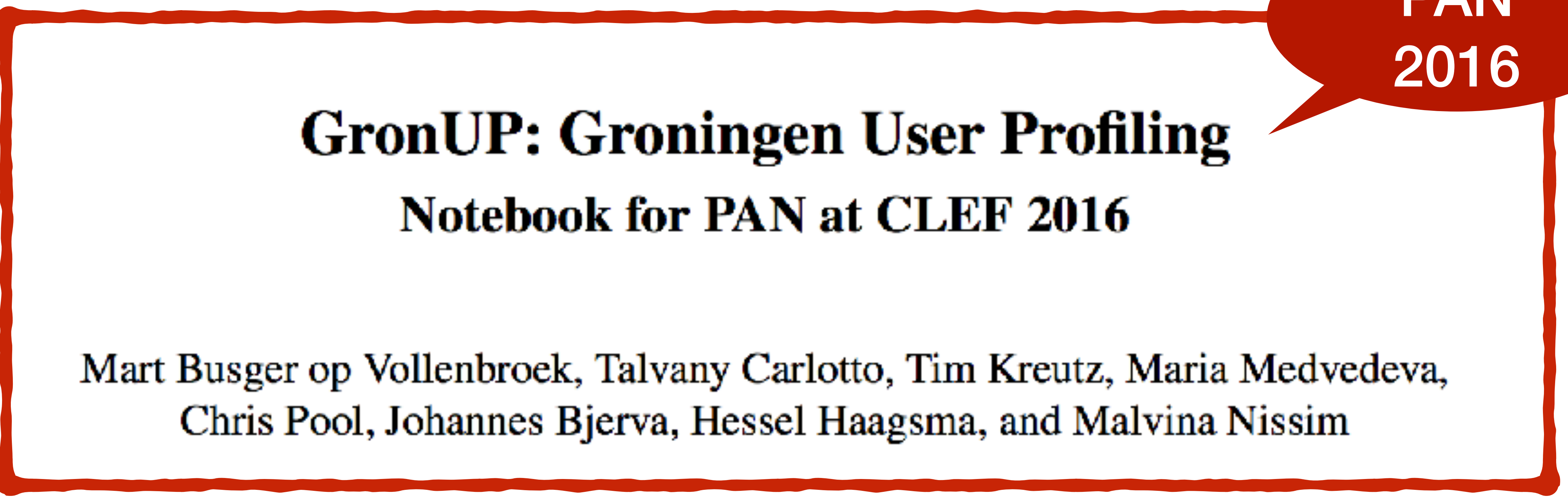
★Refine the bleaching features

★Try other lexical-poor approaches

★Expand the human experiments

★Study the results of the Cross-genre gender detection shared tasks!

★Anything you might suggest (please do!)
(this includes suggesting that we should all stop run word/chr n-gram based SVMs)

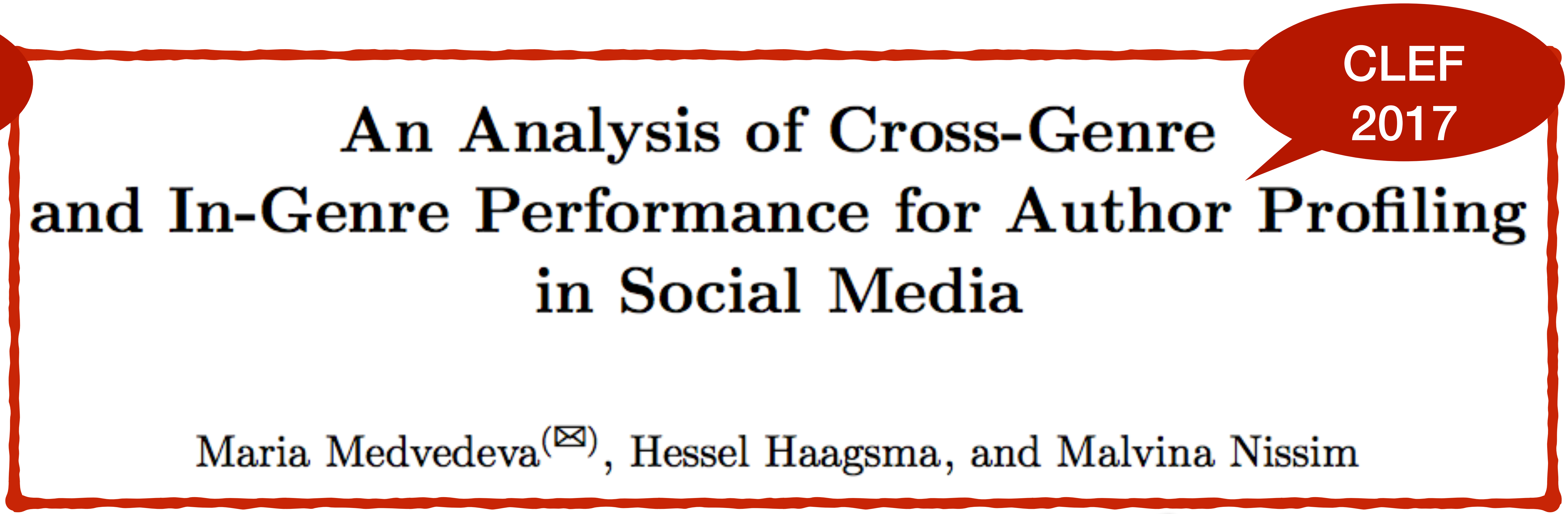


PAN 2016

GronUP: Groningen User Profiling

Notebook for PAN at CLEF 2016

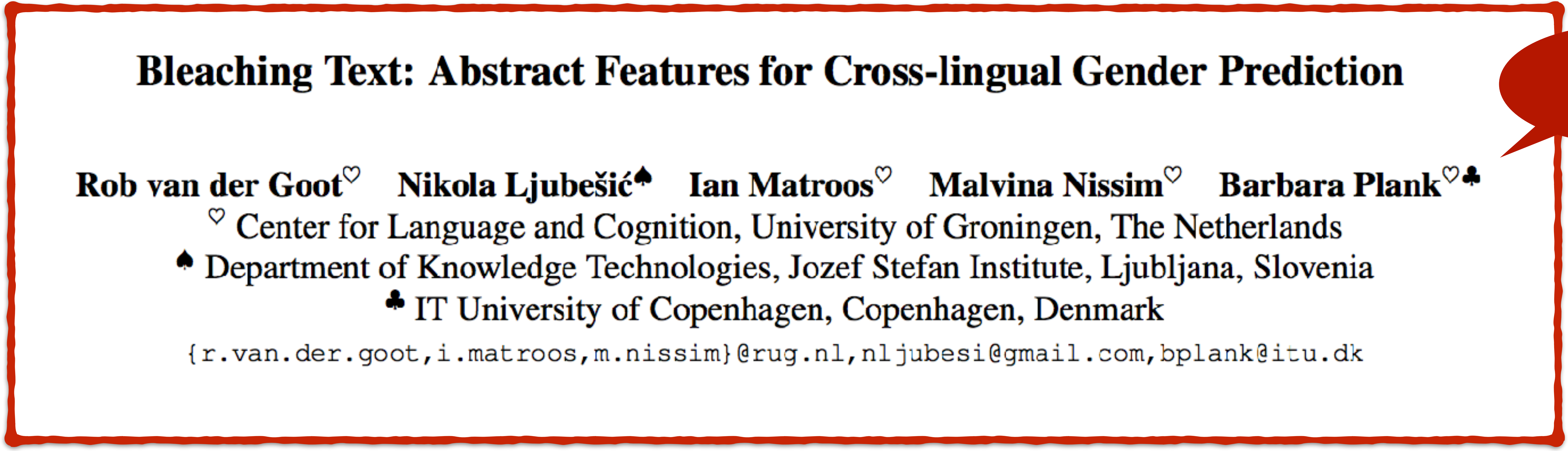
Mart Busger op Vollenbroek, Talvany Carlotto, Tim Kreutz, Maria Medvedeva, Chris Pool, Johannes Bjerva, Hessel Haagsma, and Malvina Nissim



CLEF 2017

An Analysis of Cross-Genre and In-Genre Performance for Author Profiling in Social Media

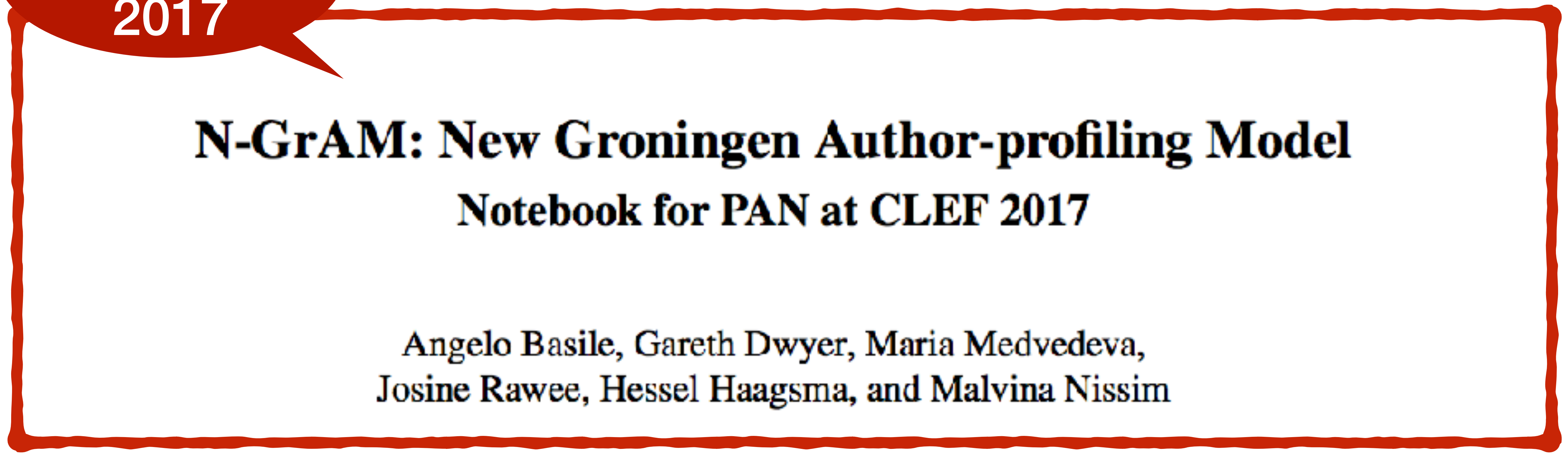
Maria Medvedeva[✉], Hessel Haagsma, and Malvina Nissim



ACL 2018

Bleaching Text: Abstract Features for Cross-lingual Gender Prediction

Rob van der Goot[♡] Nikola Ljubešić[♠] Ian Matroos[♡] Malvina Nissim[♡] Barbara Plank[♡♣]
[♡] Center for Language and Cognition, University of Groningen, The Netherlands
[♠] Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia
[♣] IT University of Copenhagen, Copenhagen, Denmark
{r.van.der.goot,i.matroos,m.nissim}@rug.nl,nljubesi@gmail.com,bplank@itu.dk

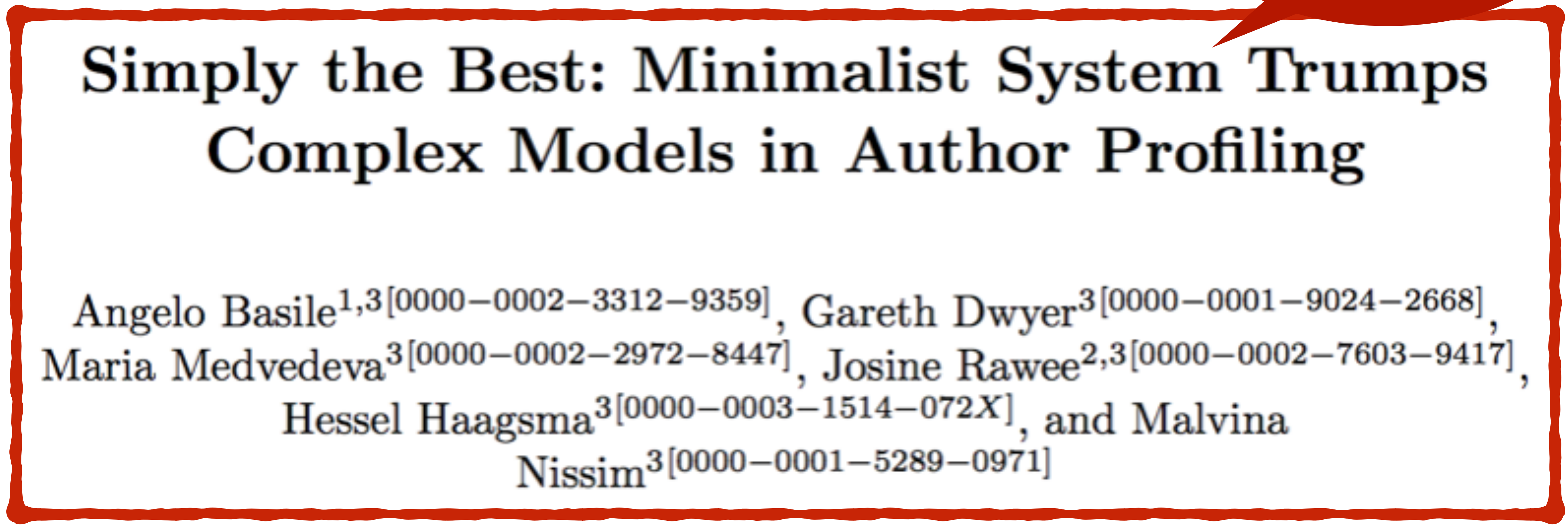


PAN 2017

N-GrAM: New Groningen Author-profiling Model

Notebook for PAN at CLEF 2017

Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim



CLEF 2018

Simply the Best: Minimalist System Trumps Complex Models in Author Profiling

Angelo Basile[1,3][0000–0002–3312–9359], Gareth Dwyer[3][0000–0001–9024–2668], Maria Medvedeva[3][0000–0002–2972–8447], Josine Rawee[2,3][0000–0002–7603–9417], Hessel Haagsma[3][0000–0003–1514–072X], and Malvina Nissim[3][0000–0001–5289–0971]