Age and Gender Identification in Social Media

James Marquardt
jamarq@uw.edu
University of Washington, Tacoma

Golnoosh Farnadi
golnoosh.farnadi@ugent.be
Ghent University

Gayathri Vasudevan
gvasu@uw.edu
University of Washington, Tacoma

Marie-Francine Moens
sien.moens@cs.kuleuven.be
Katholieke Universiteit Leuven

Sergio Davalos
sergiod@uw.edu
University of Washington, Tacoma

Ankur Teredesai
ankurt@uw.edu
University of Washington, Tacoma

Martine De Cock
mdecock@uw.edu
University of Washington, Tacoma

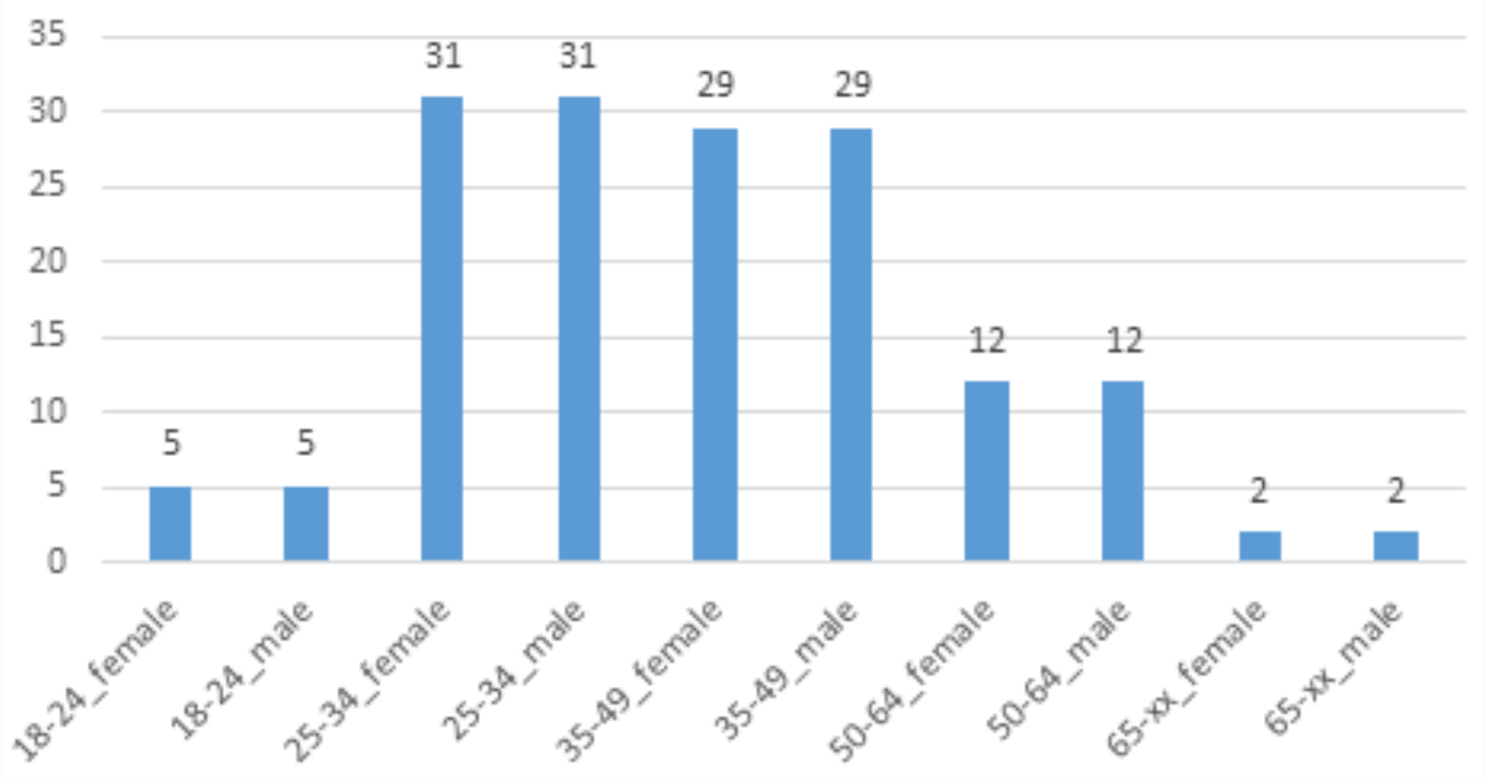
Objective / Motivation

- Construct a multi-label classification model for inferring the age and gender of the authors of text documents
- Useful for law enforcement, online reputation management, and targeted advertising

Dataset

	Document Count	
Genre	English	Spanish
Blogs	147	88
Twitter	306	178
Social Media	7,746	1272
Reviews	4,160	-

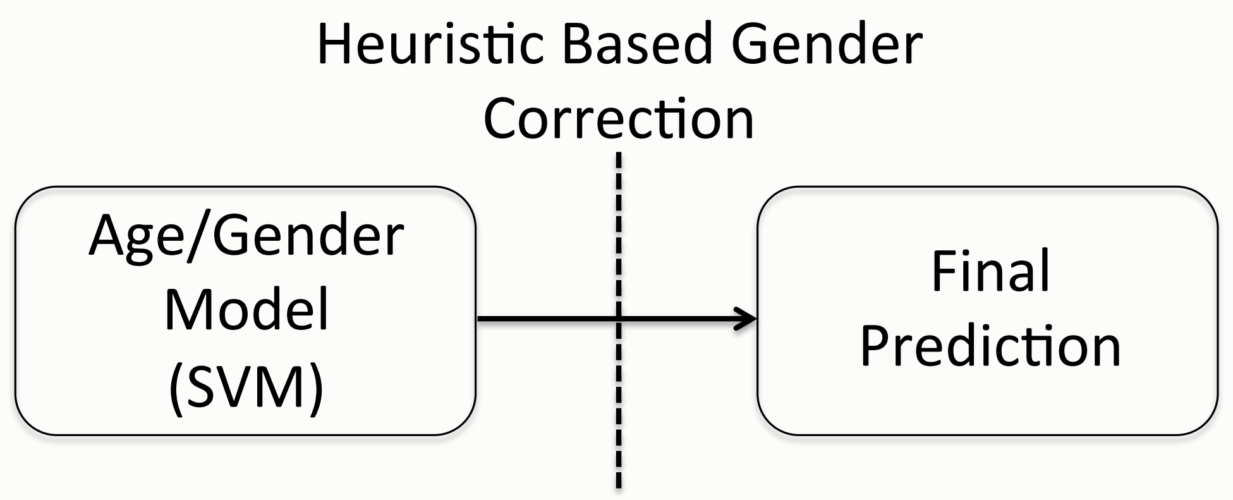
Documents from all corpora



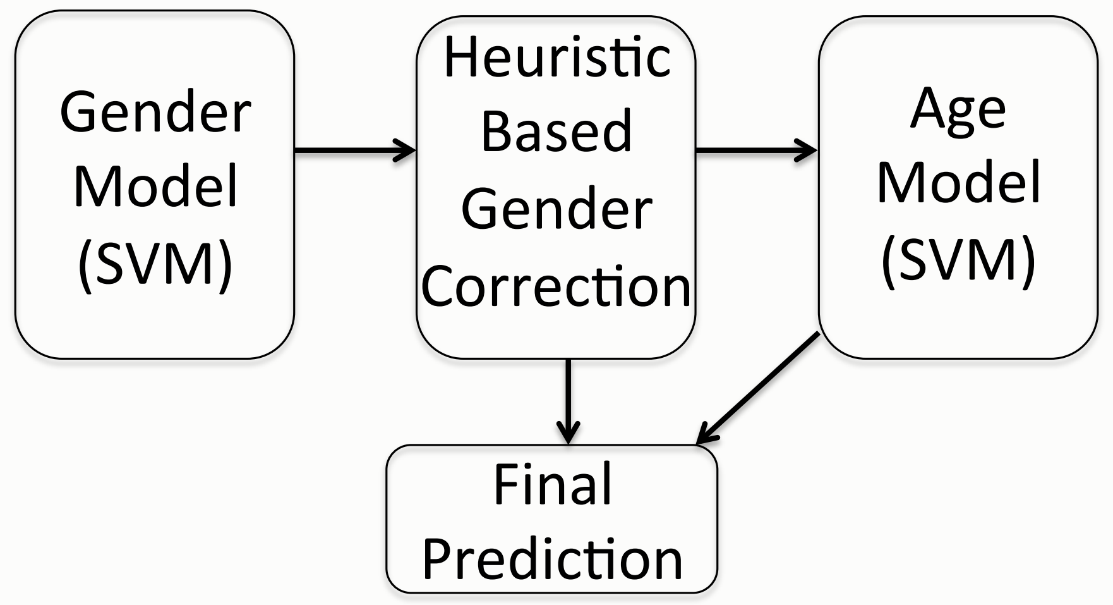
Distribution of age/gender labels in data was highly imbalanced, as shown in the English blogs corpus

Approach

- We evaluate two approaches to multi-label classification: Label Powerset Transformation and Chained Classifiers
- For both approaches, we apply heuristic based corrections after initial gender classification is done



Label Powerset Model



Chaining Classifiers

Feature Set

Feature Type	Count	Description / Source
MRC Features*	14	MRC psycholinguistic database
LIWC Words	68	Linguistic Inquiry and Word Count lexicon
Sentiment	3	SentiStrength
Readability	6	Various evaluation indexes (CLI, ARI, etc.)
HTML Tags	5	Links, images, bold, italics, lists
Spelling	2	jLanguageTool
Emoticons	1	:), ==),?(

* The MRC database contains data on psychological measures such as imagery and movement conveyed by words in the lexicon

Results

		English			Spanish		
Model	Genre	Total	Gender	Age	Total	Gender	Age
Baseline	Blogs	19.60	50.00	40.80	23.80	50.00	47.70
	Twitter	28.75	50.00	42.50	24.15	50.00	48.30
	Social Media	14.49	50.00	29.00	16.74	50.00	33.50
	Reviews	12.01	50.00	24.00	-	-	-
Label Powerset Model	Blogs	**23.12**	**68.71**	39.46	37.50	**80.68**	47.72
	Twitter	32.79	**71.15**	**46.89**	**33.71**	**74.72**	**48.31**
	Social Media	19.86	54.22	36.56	**26.10**	**64.62**	41.67
	Reviews	19.09	**65.46**	**29.83**	-	-	-
Chaining Classifiers	Blogs	23.05	66.59	**42.86**	**38.71**	72.93	**47.73**
	Twitter	**33.44**	69.15	47.73	31.62	71.35	**48.31**
	Social Media	**20.16**	**57.39**	**36.78**	24.48	63.14	**41.75**
	Reviews	**19.25**	63.11	**29.83**	-	-	-

Conclusions

- Both models outperform the baseline across multiple genres of social media
- Our models achieve slightly higher accuracies for the Spanish dataset than for the English dataset
- The benefit, in terms of accuracy, of using more the more complex chaining classifier is negligible

Acknowledgements

- This project is part of the SBO Program of the IWT-SBO-Nr. 110067.

CENTER FOR DATA SCIENCE
UNIVERSITY of WASHINGTON | TACOMA
Institute of Technology

UNIVERSITEIT GENT

KU LEUVEN

cwds.uw.edu