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

Maria Medvedeva, Hessel Haagsma, and Malvina Nissim

`hessel.haagsma@rug.nl`

Centre for Language and Cognition, University of Groningen

CLEF 2017
Dublin, 12 September 2017

Author profiling @ PAN 2016

Author Profiling: given an author's texts, predict their characteristics

- Training documents of one genre (e.g. Twitter, blogs, social media...)
- Evaluation on another genre (e.g. reviews)
- Train on Twitter, test on Blogs (unknown)
- Gender: male/female, 50/50 distribution
- Age: 18–24/25–34/35–45/50–64/65+, unequal distribution

Results

Team	Average Joint Accuracy
Busger op Vollenbroek et al. (GronUP)	0.5258
Modaresi et al.	0.5247
Bilan et al.	0.4834

GronUP

Issue

- Unknown cross-genre test set

Design Intentions

- Keep it simple **(!)**
- Ignore tweet-specific traits

Classifier

- Tweet-to-sentence preprocessing
- SVM with linear kernel and default parameters (scikit-learn)

Features

- Character, word, and PoS n-grams
- Word and sentence length
- Orthography and correctness
- Vocabulary

Details? See Busger op Vollenbroek et al, 2016. GronUP: Groningen User Profiling. In Working Notes of CLEF 2016 - Digital Text Forensics (PAN), pp. 846—857.

Follow-up: were we lucky?

Q1 Is the GronUP model truly cross-genre so that good results can be observed in datasets other than the PAN 2016 test set?

Q2 If the features truly capture some general aspects of demographics, can the model be trained on datasets other than Twitter and still yield a good performance?

Follow-up: were we lucky?

Q1 Is the the GronUP model truly cross-genre so that good results can be observed in datasets other than the PAN 2016 test set?

Evaluate the Twitter-trained GronUP model on other datasets and other genres, compare to PAN 2016 results and in-genre GronUP

Q1: Experiments
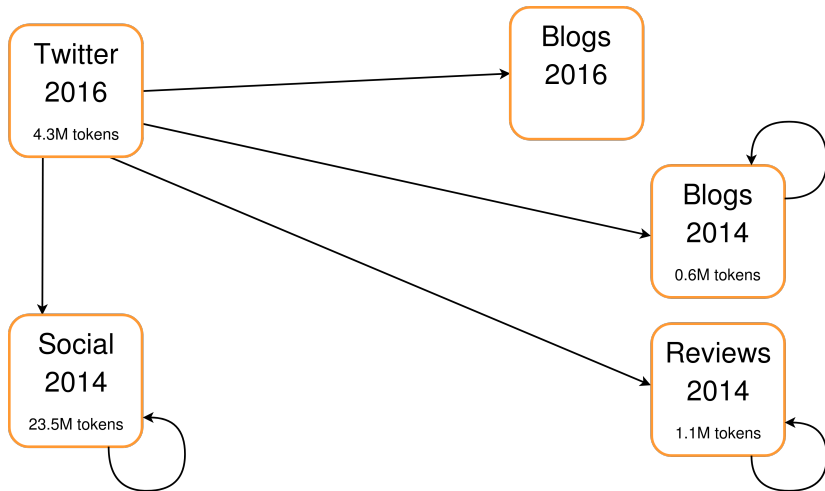
Q1: Results

Q1 Is GronUP truly cross-genre so that good results can be observed in datasets other than the PAN 2016 test set?

Training	Test	Average Acc.
Twitter 2016	Blogs 2016 (test)	0.6157
Twitter 2016	Blogs 2014	0.5936

Q1: Results

Q1 Is GronUP truly cross-genre so that good results can be observed in datasets other than the PAN 2016 test set?

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

Q1: Results

Q1 Is GronUP truly cross-genre so that good results can be observed in datasets other than the PAN 2016 test set?

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

Q1: Answer

Is GronUP truly cross-genre so that good results can be observed in
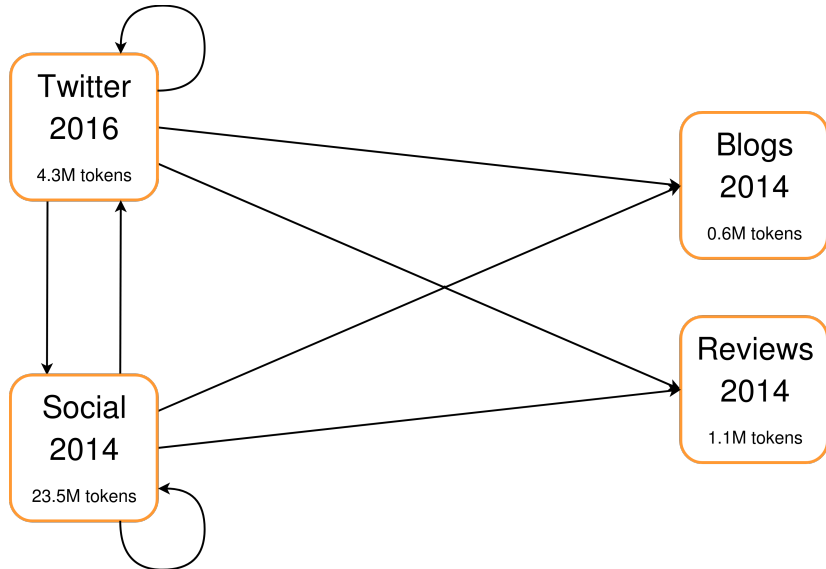datasets other than the PAN 2016 test set?

A1 **Maybe**

Follow-up: were we lucky?

Q2 If the features truly capture some general aspects of demographics, can the model be trained on datasets other than Twitter and still yield a good performance?

Retrain the GronUP model on a different genre, and compare results to the Twitter-trained model

Q2: Experiments

Q2: Results

Q2 If the features truly capture some general aspects of demographics, can the model be trained on datasets other than Twitter and still yield a good performance?

Training	Test	Average Acc.
Twitter 2016	Blogs 2014	0.5936
Twitter 2016	Reviews 2014	0.3689
Twitter 2016	Social Media 2014	0.4240

Q2: Results

Q2 If the features truly capture some general aspects of demographics, can the model be trained on datasets other than Twitter and still yield a good performance?

Training	Test	Average Acc.
Twitter 2016	Blogs 2014	0.5936
Twitter 2016	Reviews 2014	0.3689
Twitter 2016	Social Media 2014	0.4240
Social Media 2014	Blogs 2014	0.5045
Social Media 2014	Reviews 2014	0.3764
Social Media 2014	Twitter 2016	0.4535

Q2: Results

Q2 If the features truly capture some general aspects of demographics, can the model be trained on datasets other than Twitter and still yield a good performance?

Training	Test	Average Acc.
Twitter 2016	Blogs 2014	0.5936
Twitter 2016	Reviews 2014	0.3689
Twitter 2016	Social Media 2014	0.4240
Social Media 2014	Blogs 2014	0.5045
Social Media 2014	Reviews 2014	0.3764
Social Media 2014	Twitter 2016	0.4535
Twitter 2016	Cross-Validation	0.5906

Q2: Answer

Q2 If the features truly capture some general aspects of demographics, can the model be trained on datasets other than Twitter and still yield a good performance?

A2 **Maybe**

Conclusions

- Size matters
- Genre matters
- Cross-genre is feasible and valuable

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

The End