CAPS: A Cross-genre Author Profiling System

Ivan Bilan and Desislava Zhekova
Center for Information and Language Processing, LMU Munich, Germany
ivan.bilan@gmx.de
zhekova@cis.uni-muenchen.de

Presentation Overview

- Overview of Author Profiling
- Training Dataset
- Software Tools
- Machine Learning Pipeline
- Custom Features
- Classification
- Final Results

Overview of Author Profiling

Author Profiling – attributing an author of a text to a certain sociodemographic class

Real world applications:

» suspect profiling in forensics

» customer-base analysis

» targeted advertising

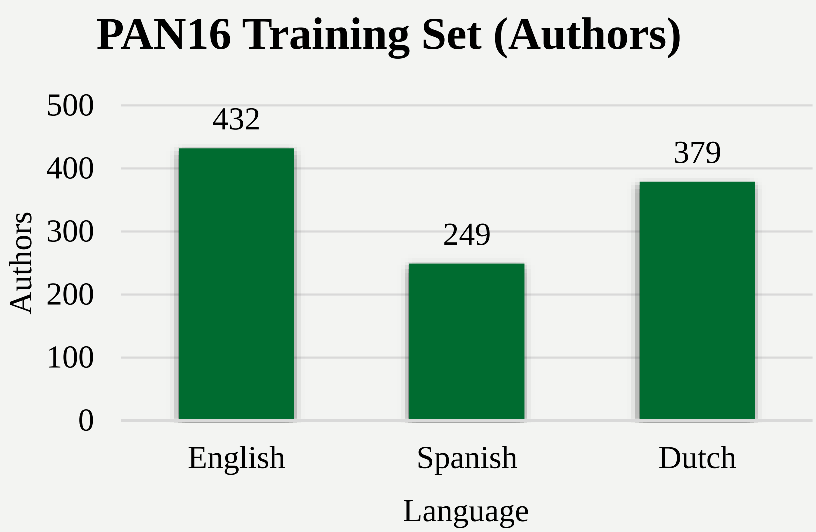
Cross-genre author profiling:

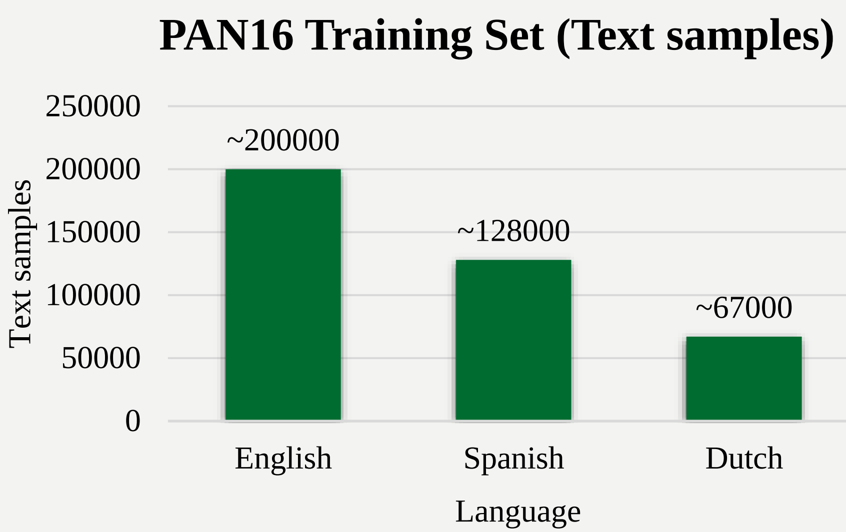
» adaptable to any unseen genre

» label only genres that are easier to label

» merge all existing genres into one training set to overcome data scarcity

Training Dataset





» <u>Labelled with gender</u>: Male Female

» <u>Age groups</u>: 18-24 25-34 35-49 50-64 65-xx

» Artificially increase the number of samples by labeling each text sample

» During evaluation take the most frequent prediction (or the one with the highest confidence score) for the author

Software Tools

» Python

» scikit-learn (main machine learning toolkit)

» gensim (topic modelling)

» matplotlib (visualization)

» TreeTagger (available at http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/)

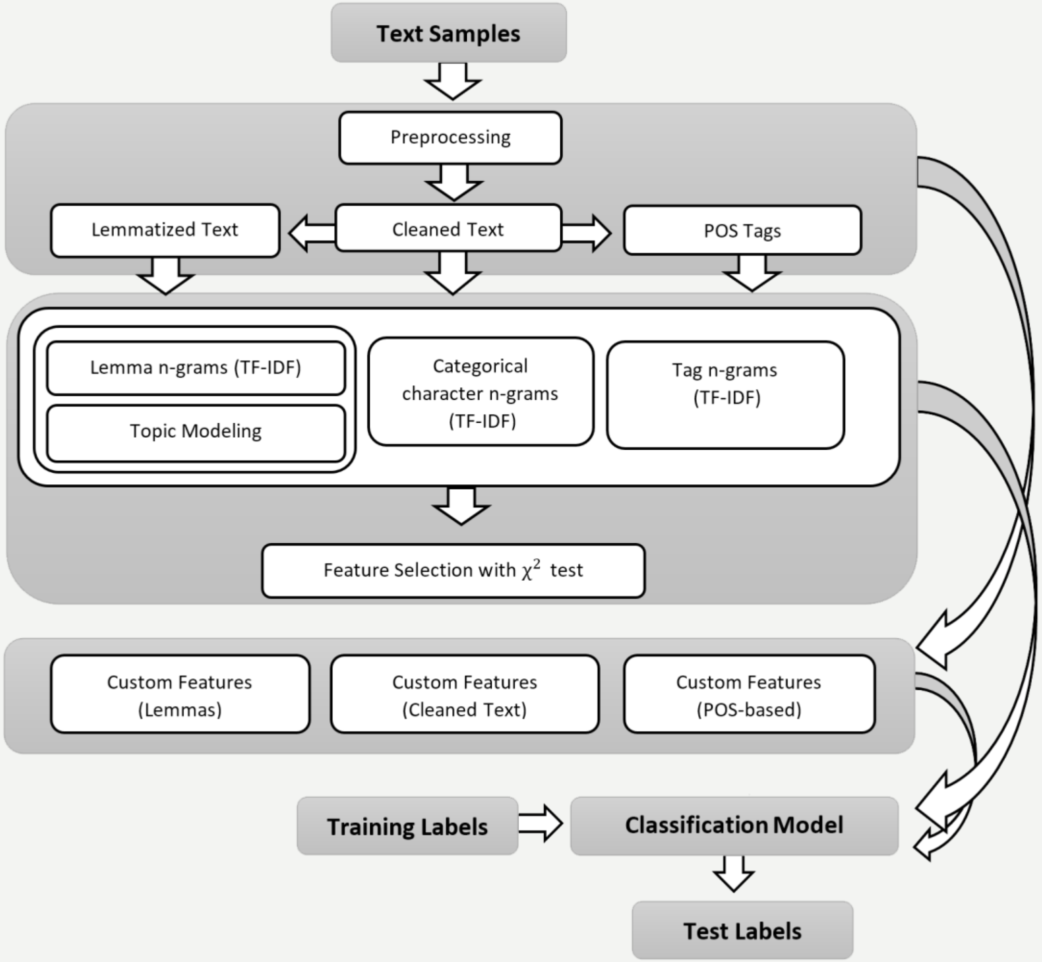
  » supports part-of-speech tagging, lemmatization, stemming and chunking

  » works on multiple languages

  » has wrappers for various programming languages

  » freely available for research and education

Machine Learning Pipeline



Machine Learning Pipeline

Preprocessing

- HTML and Bulletin Board Code removal
- normalization of all links to *[URL]*
- normalization of all usernames e.g. @username to *[USER]*
- duplicate sample removal

Text representations

- first experimented with stemmed text representation
- final system uses lemma and part-of-speech representation
- the results are saved in a dataframe and each feature accesses the text representation it requires

Machine Learning Pipeline

TF-IDF - The Term Frequency-Inverse Document Frequency

- » Emphasize important words (frequent in a text, infrequent in the corpus)

Usage in CAPS:

- » unigrams, bigrams, trigrams for lemmatized text
- » 1-4 grams for POS text representation
- » 3-grams for characters

Topic Modelling with Latent Dirichlet Allocation (LDA)

and Hierarchical Dirichlet Process (HDP)

- » Generative statistical model that allows automated grouping of observed words into topics
- » LDA requires predefined number of topics
- » HDP calculates the number of topics automatically
- » do not confuse with linear discriminant analysis (also known as LDA)

Usage in CAPS:

- » we used LDA with 100 topics
- » HDP showed decreased performance

Custom Features

- Over 40 custom features divided into the following feature clusters:
  - Dictionary-based Features
  - POS-Based Features
  - Text Structure Features
  - Stylistic Features

Dictionary-based Features

Feature Cluster		Examples per Language		
	Feature Name	**English**	**Spanish**	**Dutch**
Dictionary-based	Connective Words	*furthermore, firstly …*	*pues, como …*	*zoals, mits …*
	Emotion Words	*sad, bored, angry …*	*espanto, carino, calma …*	*boos, moe, zielig …*
	Contractions	*I'd, let's, I'll …*	*al, del, desto …*	*m'n, 't, zo'n …*
	Familial Words	*wife, husband, gf …*	*esposa, esposo …*	*vriendin, man …*
	Collocations	*dodgy, awesome, troll …*	*no manches, chido …*	*buffelen, geil …*
	Abbreviations and Acronyms	*a.m., Inc., asap …*	*art., arch. …*	*gesch., geb. …*
	Stop Words	*did, we, ours …*	*de, en, que …*	*van, dat, die …*

» positive / negative sentiment lists are not used

POS-Based Features

- » Use of Verbs, Interjections, Adjectives, Determiner, Conjunction, Plural Nouns
- » Lexical Measure – tell how implicit or explicit the text is

$$F = 0.5\left(\left((nouns + adjectives + prepositions + articles) - (pronouns + verbs + adverbs + interjections)\right) + 100\right)$$

Heylighen et al. (2002)

Readability Index Formulas

- » tried Automated Readability Index, SMOG Readability Formula, Flesch Reading Ease etc.
- » decreased effectiveness in cross-genre setting since
- » not suitable for short text samples
- » e.g. Flesch Reading Ease: $206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right)$

Text Structure Features

- Type/Token ratio
- Average word length
- Usage of punctuation marks

Stylistic features (occurrence of adjectival endings)

- English: *-ly, -able, -ic, -il, -less, -ous etc.*
- Spanish: *-ito, -ada, -anza, -acho, -acha etc.*
- Dutch: *-jes, -iek, -eren etc.*

Feature Scaling

Step 1: Scale to sample length

- the feature vector values are divided by the sample length

$$x^{(i)}_{pre-scaled} = \frac{feature\ vector\ value}{len(sample)}$$

Step 2: Standardize

$$x^{(i)}_{std} = \frac{x^{(i)}_{pre-scaled} - \mu_x}{\sigma_x}$$

- $x^{(i)}_{pre-scaled}$ is a feature vector sample
- μ_x is sample mean of the feature column
- σ_x represents the standard deviation of the feature column

Classification

Gender and age classified separately:

- Support Vector Machine (namely Linear Support Vector Classification) classifier used for gender classification
- Multinomial Logistic Regression for age classification

Final Results (Cross-genre)

PAN16 Results, Accuracy (Cross-genre, all represented languages)

PAN16	English			Spanish			Dutch
Class	Gender	Age	Both	Gender	Age	Both	Gender
Best Score	75.64%	58.97%	39.74%	73.21%	51.79%	42.87%	61.80%
CAPS	74.36%	44.87%	33.33%	62.50%	46.43%	37.50%	55.00%
Lowest Score	46.15%	32.05%	14.10%	46.43%	21.43%	21.43%	41.60%

Final Top 5 Ranking (PAN16, by overall average)

Place:	1st	2nd	3rd (CAPS)	4th	5th
Result:	52.58%	52.47%	48.34%	46.02%	45.93%

Final Results (Single genre)

» the system also performs rather effectively in single genre setting

PAN14 and PAN15 Results, Accuracy (Single genre, English)

PAN14-15	Twitter (PAN15)		Blogs (PAN14)		Hotel Reviews (PAN14)	
Class	**Gender**	**Age**	**Gender**	**Age**	**Gender**	**Age**
Best Score	85.92%	83.80%	67.95%	46.15%	72.59%	35.02%
CAPS	81.69%	73.24%	66.67%	35.90%	71.32%	34.77%

Future work

- use dependancy parsing and extract features based on the tree representation
- improve features for Spanish and Dutch

Thank you for your attention!

References

1. Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, (pp. 44-49). Manchester, UK.
2. Spärck Jones, K. (1972). A Statistical Interpretation of Term Specificity and its Retrieval. Journal of Documentation, 28(1), 11-21.
3. Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. The Journal of Machine Learning Research, 3(1), 993-1022.
4. Heylighen, F., Dewaele, J.: Variation in the Contextuality of Language: An Empirical Measure. Foundations of Science 7(3), 293–340 (2002)
5. Flesch, F. (1948). A new readability yardstick. The Journal of applied psychology, 32(3), 221-233.