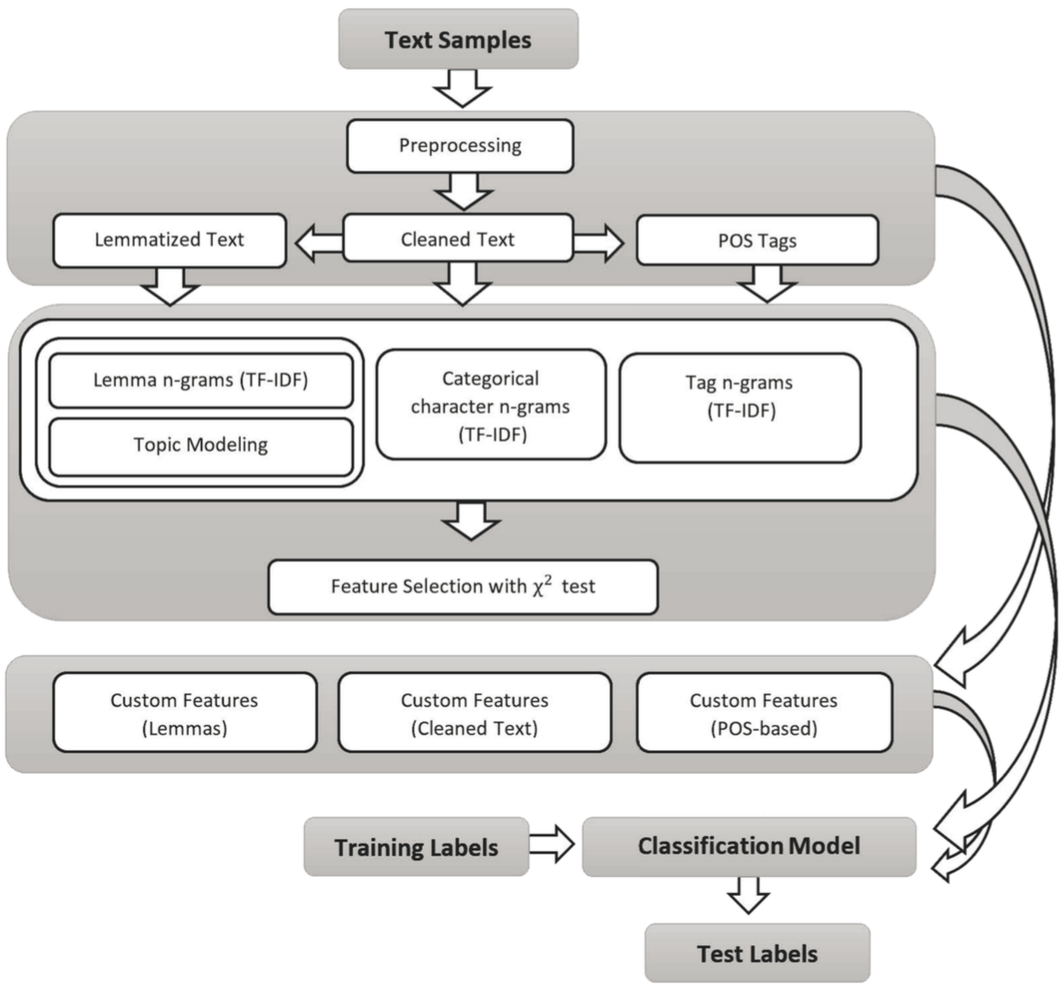
CAPS: A Cross-genre Author Profiling System

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

1. CAPS Overview



2. Preprocessing

- HTML, Bulletin Board Code removal
- normalization of Links ([URL]), Usernames e.g. @username ([USER])
- lemma and POS annotation via the TreeTagger (Schmid, 1994)
- duplicate sample removal:

Language		Text Samples					Unique Authors				
English	***Age***	*18-24*	*25-34*	*35-49*	*50-64*	*65-xx*	*18-24*	*25-34*	*35-49*	*50-64*	*65-xx*
	Samples	15725	68936	79338	34668	1435	28	137	181	80	6
	Gender	*Male*			*Female*		*Male*			*Female*	
	Samples	111030			89072		216			216	
	Total	*200102 Text Examples*					*432 Authors*				
Spanish	***Age***	*18-24*	*25-34*	*35-49*	*50-64*	*65-xx*	*18-24*	*25-34*	*35-49*	*50-64*	*65-xx*
	Samples	7146	30730	66287	21449	2869	16	63	38	20	6
	Gender	*Male*			*Female*		*Male*			*Female*	
	Samples	70129			58352		124			125	
	Total	*128481 Text Examples*					*249 Authors*				
Dutch	***Gender***	*Male*			*Female*		*Male*			*Female*	
	Samples	33111			33773		188			191	
	Total	*66884 Text Examples*					*379 Authors*				

3. Feature Extraction

- TF-IDF lemmas (uni-, bi- and trigrams), POS-tags (uni-, bi-, tri-, fourgrams), characters (trigrams)
- LDA topic modelling - 100 different topics
- dictionary-based (connective words, emotion words, contractions, family related words, collocations, abbreviations, acronyms, stop words);
- POS-based – use of verbs, interjections, adjectives, determiners, conjunctions, plural nouns, lexical measure; includes a more complex F-Measure feature following Heylighen et al. (2002)
- text structure – e.g. type/token ratio, average word length, use of punctuation marks
- stylistic – frequency of use of adjectival and adverbial suffixes – e.g. -ly,-able,-ic,-il,-less,-ous etc.
- readability index – Automated Readability Index, SMOG Readability Formula, Flesch Reading Ease (not effective for the cross-genre setting)
- chi-square term selection for dimensionality reduction

Feature Cluster	Feature Name	Feature Value Examples		
		English	**Spanish**	**Dutch**
Dictionary-based	Connective Words	*furthermore, firstly, moreover, hence ...*	*pues, como, luego, aunque ...*	*zoals, mits, toen, zeker ...*
	Emotion Words	*sad, bored, angry, nervous, upset ...*	*espanto, carino, calma, peno ...*	*boos, moe, zielig, chagrijnig ...*
	Contractions	*I'd, let's, I'll, he'd, can't, he'd ...*	*al, del, desto, pal', della ...*	*m'n, 't, zo'n, a'dam ...*
	Familial Words	*wife, husband, gf, bf, mom ...*	*esposa, esposo, marido, amiga ...*	*vriendin, man, vriend, moeder ...*
	Collocations	*dodgy, telly, awesome, freak, troll ...*	*no manches, chido, sale ...*	*buffelen, geil, dombo, tjo ...*
	Abbreviations and Acronyms	*a.m., p.m., Mr., Inc., NASA, asap ...*	*art., arch., Avda., Arz., ant. ...*	*gesch., geb., nl, notk, mv, vnl ...*
	Stop Words	*did, we, ours, you, who, these, because ...*	*de, en, que, los, del, donde, como ...*	*van, dat, die, was, met, voor ...*

4. Feature Scaling

- The sample length is rescaled relative to the lowest mean length of a text sample throughout all possible writing styles that could be represented in both training and test sets.
- The feature values are divided by this rescaled sample length.
- The rescaled sample length represents the amount of possible smallest sample entities that would fit into the text sample under review.

$$x^{(i)}_{pre-scaled} = \frac{x^{(i)}}{\left(\frac{len(\varepsilon_i)}{min(\mu_{y_1} \dots \mu_{y_n}) \mid y_n := len(\varepsilon_{m_1}) \dots len(\varepsilon_{m_n})}\right)} \quad (1)$$

$$x^{(i)}_{std} = \frac{x^{(i)} - \mu_x}{\sigma_x} \quad (2)$$

5. Classification

- classify each text sample independent of the others
- classify the author class based on each text sample belonging to the author
- gender (LinearSVC) and age (Multinomial Logistic Regression) are also classified independently

6. Evaluation

6.1 Final PAN16 results for CAPS

Language (Setting)	CAPS			PAN16 Best	Baseline
	Test Set 1	Test Set 2	Average		
English (Gender)	53.74	74.36	64.05	75.64	56.41
English (Age)	29.02	44.87	36.95	58.97	19.23
Spanish (Gender)	56.25	62.50	59.38	73.21	50.00
Spanish (Age)	23.44	46.43	34.94	51.79	17.86
Dutch (Gender)	54.00	55.00	54.50	61.80	53.00

6.2 Results on the PAN15 Datasets

Language (Setting)	CAPS			PAN15 Best	Baseline
	Test Set 1	Test Set 2	Average		
English (Gender)	85.71	81.69	83.70	85.92	50.00
English (Age)	73.81	73.24	73.53	83.80	25.00
Spanish (Gender)	93.33	88.64	90.99	96.59	50.00
Spanish (Age)	66.67	67.05	66.86	79.55	25.00
Dutch (Gender)	80.00	78.13	79.07	96.88	50.00

6.3 Results on the PAN14 Datasets

Language (Setting)	Genre	CAPS			PAN14 Best	Baseline
		Test Set 1	Test Set 2	Average		
English (Gender)	Blogs	58.33	66.67	62.50	67.95	57.69
English (Age)	Blogs	25.00	35.90	30.45	46.15	14.10
English (Gender)	Twitter	63.33	60.39	61.86	73.38	59.74
English (Age)	Twitter	56.67	45.45	51.06	50.65	27.92
English (Gender)	Hotel Reviews	73.78	71.32	72.55	72.59	66.26
English (Age)	Hotel Reviews	37.20	34.77	35.99	35.02	27.53
Spanish (Gender)	Blogs	42.86	42.86	42.86	58.93	53.57
Spanish (Age)	Blogs	35.71	44.64	40.18	48.21	16.07
Spanish (Gender)	Twitter	61.54	56.67	59.11	65.56	47.78
Spanish (Age)	Twitter	46.15	48.89	47.52	61.11	46.67

7. Conclusion and Future Work

- CAPS was ranked third from all 22 teams that participated in PAN16.
- CAPS achieved the second best score – 74.36% accuracy (with the best performing system reaching 75.64%) for gender identification on the official PAN16 test set for English.
- CAPS also proved to be very competitive to the state-of-the-art systems on both PAN15 and PAN14 datasets.
- CAPS achieved 81.69% for gender classification on the PAN15's English dataset with the best PAN15 participating system reaching a performance of 85.92%.
- CAPS would profit from language specific features for all languages other than English, since currently the feature set is tailored to English.
- CAPS could also profit from text sample-author profile interrelation, which we plan to explore.