

Even Blogs in the Wild: Large Scale Personality Classification

Francisco Iacobelli: f-iacobelli@u.northwestern.edu
 Alastair Gill: A.Gill@surrey.ac.uk
 Scott Nowson: snowson@appen.com.au
 Jon Oberlander: J.Oberlander@ed.ac.uk

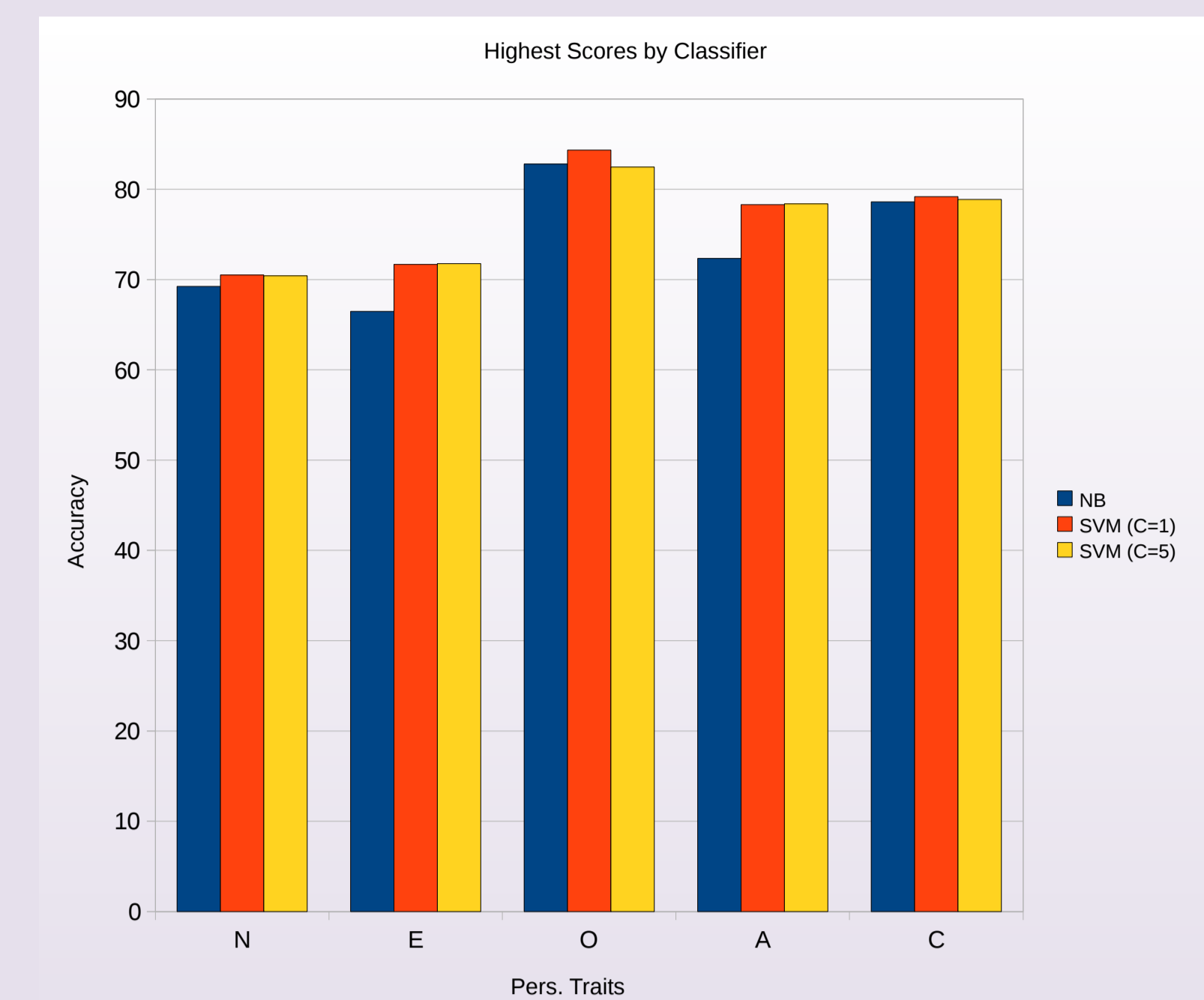
Motivation

- Personality has been linked both to reader preferences and writer motivation
- Automatic identification of personality in an online context could improve the personalisation of content
- Benefits content providers, advertisers and users.

Personality

Trait	High Score	Low Score
Neuroticism	Emotional instability; anxious; hostile; prone to depression	Emotional stability; calm; less easy upset
Extraversion	Extraverts; warm; assertive; action-oriented; thrill-seeking	Introverts; low key; deliberate; easily stimulated
Openness	Appreciate art and ideas; imaginative; aware of feelings	Straightforward interests; conservative; resist change
Agreeableness	Compassionate; cooperative; considerate; friendly	Suspicious; unfriendly; wary; antagonist; uncooperative
Conscientiousness	Disciplined; dutiful; persistent; compulsive; perfectionist	Spontaneous; impulsive; achievement less important

Which Classifier?

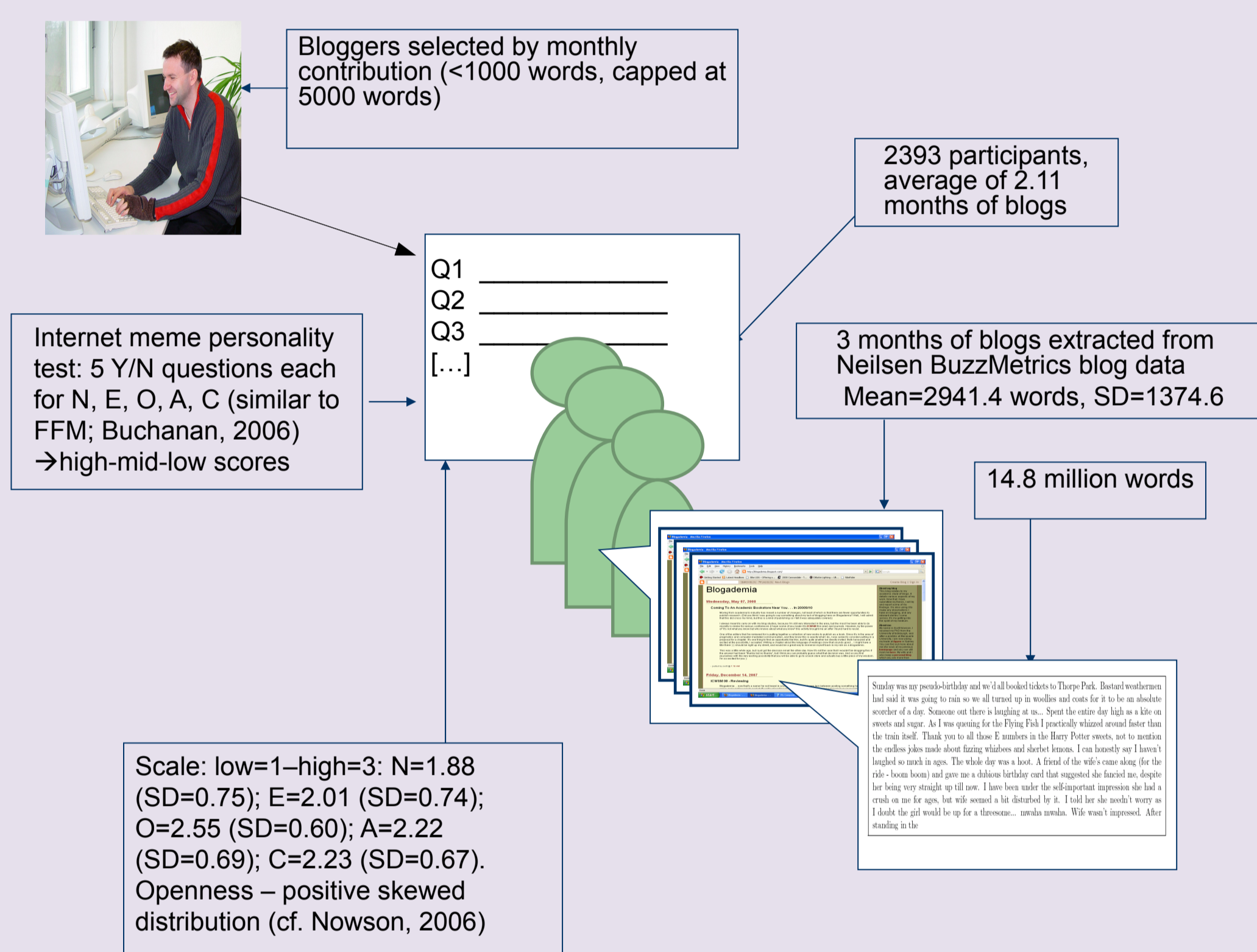


Which Featureset?

Trait	bool		freq		TAWC
	2	1	1	2	
N	70.51	70.47	70.12	67.77	59.56
E	71.68	67.80	68.40	63.99	54.86
O	84.36	81.44	79.14	77.49	56.86
A	78.31	69.98	69.49	71.09	61.09
C	79.18	75.17	72.74	76.41	56.11

Table 2. Comparison of featuresets using SVM (C=1). ◦ improvement; degradation at p<0.05 with respect to “bool;2;sw”

Data



Level	N	E	O	A	C
High	553/840	637/637	137/137	372/372	323/323
Low	553/553	637/669	137/1465	372/892	323/884
Total	1106/1393	1274/1306	274/1602	744/1264	646/1207

Table 1. Number of docs used for experiment out of total number of docs; by trait and score.

Which bigrams drive classification?

Trait	Categories	High	Low
Neuroticism	Problem Talk	Only problem depressed you	Be sad
	Exaggeration	Very low	-
	Self references	I wasn't drunk I	Am excited
	Pronouns	Put her hope they you only	-
	Thoughtful	-	Reflect on choose to
	Inferred Claims	-	Then look great because
Extraversion	Tentative	-	But other
	Strong curse	You f**k b**ch I what f**k	-
	Informality	I'm happy I'm at weird on	Wait so
	Self references	I miss dance I love me	Increase my my regular
	Long words	-	Favorite character
	Tentative	-	Intend to
Openness	Prepositions	-	Put away
	Weak cursing	The hell like sh*t	-
	Proper names	All <proper name (NNP)>	Monday and
	Self references	I lost pick me	-
	Categories	-	You belong
	Religion	-	To church at church pray for
Agreeableness	disagreement	-	Not exactly
	Positive words	Even better of beauty	-
	Disc. Markers	Doh oh really all me sigh	-
	Self references	You I'll keep myself	Self interest case I along I
	negativity	-	Unfortunately the what worry
	Conscientiousness	Plan and eval.	To study on track prior to
Desire		Hopefully I hope I'm	-
Justifications		-	Real reason of why
Self references		I work test I I work year I	How I'm am also

Table 3. Categorization of bigrams per personality trait. This is not a comprehensive list. Just an illustration of the data organized by categories from previous research.

Features

Word scoring:

Boolean (1 or 0 for the presence or absence of a word)

Frequency (score derived from TF-IDF);

Window size of extracted features (1-grams vs 2-grams);

Inclusion (sw) or **omission (wo)** of stop words.

Augmented LIWC categories as implemented by TAWC

(Pennebaker and Francis, 1999; Kramer, Oh and Fussell, 2006).

Conclusion

• Best classifier: SVM (C=1). Best accuracy for Openness to Experience (84.36%) and worst for Neuroticism (70.51%).

• Best accuracies were using:

Bigrams: So language structure matters, at least to some degree.

Stop words: They seem to matter in personality expression. This fits with previous observations that (for instance) even punctuation reflects personality.

• Categories of words may be narrower than previously thought. (degrees of curse words, exaggeration, etc.)

Degrees may tip classification.

TAWC: performance may be due to its grouping of words that shouldn't be in the same category for the purposes of pers. classification.

