Effects of Age and Gender on Blogging

Jonathan Schler [1] **Moshe Koppel** [1] **Shlomo Argamon** [2] **James Pennebaker** [3]

[1] Dept. of Computer Science, Bar-Ilan University, Ramat Gan 52900,Israel [2] Linguistic Cognition Lab, Dept. of Computer Science Illinois Institute of Technology, Chicago, IL 60616 [3] Dept. of Psychology, The University of Texas, Austin, TX 78712

schlerj@cs.biu.ac.il, koppel@cs.biu.ac.il, argamon@iit.edu, pennebaker@mail.utexas.edu

Abstract

Analysis of a corpus of tens of thousands of blogs – incorporating close to 300 million words – indicates significant differences in writing style and content between male and female bloggers as well as among authors of different ages. Such differences can be exploited to determine an unknown author's age and gender on the basis of a blog's vocabulary.

Introduction

The increasing popularity of publicly accessible blogs offers an unprecedented opportunity to harvest information from texts authored by hundreds of thousands of different authors. Conveniently, many of these blogs include formatted demographic information provided by the authors. (While much of this information is no doubt spurious, it is reasonable to assume that most is not.) Moreover, the blog genre imposes no restrictions on choice of topic. Thus, we can exploit this rich data set to begin to answer the following question: How do content and writing style vary between male and female bloggers and among bloggers of different ages? How much information can we learn about somebody simply by reading a text that they have authored? These are very basic questions that are both of fundamental theoretical interest and of great practical consequence in forensic and commercial domains.

In the following sections, we will describe our corpus, highlight the significant differences in writing style and content among different populations, and show the extent to which such differences can be exploited to determine an unknown author's age and gender. These differences confirm findings reported earlier with regard to age [8] and gender [1,5,7] on the basis of much more limited corpora, and introduce many new findings.

The Corpus

For the purpose of this study, we reviewed all blogs accessible from blogger.com one day in August 2004, downloading each one that included author-provided indication of gender and at least 200 appearances of common English words. The full corpus thus obtained included over 71,000 blogs.

Each downloaded blog includes all blog entries from inception to the harvesting date. Formatting within a blog entry was ignored; in particular, we did not distinguish between quotes within a blog and the text of the blog itself. Some of these blogs include additional text in languages other than English; this text is ignored for our purposes. Some of the blogs included author-provided indication of age.

Figure 1 shows the distribution of blogs over various categories for age and gender. Given the large number of blogs in our corpus, it is reasonable to assume that our corpus reflects the actual distribution of all blogs for these attributes. Note that among bloggers up to the age of 17, females are a distinct majority (63%), but among bloggers above the age of 17, females are a minority (47%).

	gender		
age	**female**	**male**	**Total**
unknown	12287	12259	24546
13-17	6949	4120	11069
18-22	7393	7690	15083
23-27	4043	6062	10105
28-32	1686	3057	4743
33-37	860	1827	2687
38-42	374	819	1193
43-48	263	584	847
>48	314	906	1220
Total	34169	37324	71493

Table 1 Blogs Distribution over Age and Gender

To prevent bias, we created a sub-corpus consisting of an equal number of male and female blogs in each age group, by randomly discarding surplus documents in the larger category. This left us with a total of 37,478 blogs (the sum of the minima over the various age brackets), comprising 1,405,209 blog entries and 295,526,889 words. All statistics reported in this paper are taken from this sub-corpus.

DISTINGUISHING FEATURES

In this section we consider differences in male and female blogging and differences among bloggers of different ages. Broadly speaking, two different kinds of potential distinguishing features can be considered: style-related and content-related.

For our purposes, we consider three types of style-related features: selected parts-of-speech, function words and blog-specific features such as "blog words" and hyperlinks. (Blog words are neologisms – such as *lol*, *haha* and *ur* – that appear with high frequency in the blog corpus.)

Our content-related features are simple content words, as well as special classes of words taken from the handcrafted LIWC [9] categories, as will be described below.

Style-based features

For each of these features, we measured the frequency with which it appears in the corpus per gender per age bracket. For simplicity of presentation, we show (Table 2 – see end of paper) frequencies per gender and age bracket (13-17 [=10s]; 23-27 [=20s]; 33-42 [=30s]) of a variety of parts of speech as well as of hyperlinks and blog words. Average number of words per post is also shown.

First, note that for each age bracket, female bloggers use more pronouns and assent/negation words while male bloggers use more articles and prepositions. Also, female bloggers use blog words far more than do male bloggers, while male bloggers use more hyperlinks than do female bloggers. All of this confirms and extends findings reported earlier in [1,5,7] and lends support to the hypothesis that female writing tends to emphasize what Biber [3] calls "involvedness", while male writing tends to emphasize "information".

Another point to note is that prepositions and articles, which are used more frequently by male bloggers, are used with increasing frequency by all bloggers as they get older. Conversely, pronouns, assent/negation words and blog words, which are used more frequently by female bloggers, are used with decreasing frequency as bloggers get older. In short, the very same features that distinguish between male and female blogging style also distinguish between older and younger blogging style. (Recall that the corpus is entirely balanced for gender within each age bracket.)

Content-based features

Gender: In Table 3, we show frequency (per 10,000 words) and standard error – among male and female bloggers, respectively – for those content-based unigrams with greatest information gain for gender. The features shown are the 15 words that appear at least 5000 times in the corpus and have highest information gain among those that appear more frequently among males and females, respectively.

The differences shown here are all significant at $p<.00001$. These differences further suggest a pattern of more "personal" writing by female bloggers than male bloggers. This is further borne out in Table 4, where we show the same statistics for classes of words taken from LIWC classes.

Note that LIWC does not include word classes relating to politics and technology. However, as indicated by Table 3 – and as confirmed by perusal of the full frequency tables (too large to be shown here) – male blogging is characterized by far more references to politics and technology.

Age: In Table 5, we show we show frequency (per 10,000 words) and standard error – among bloggers in their 10s, 20s and 30s, respectively – for those content-based unigrams with greatest information gain for age. The features shown are the 10 words that appear at least 5000 times in the corpus and have highest information gain among those that appear more frequently among 10s, 20s and 30s, respectively.

In Table 6, we show the same statistics for the same LIWC word classes considered above.

Taken together, Tables 5 and 6 tell the story of the lives of the class of people who have the resources to write blogs. Teenage concern with friends and mood swings gives way to the indulgences of college life and then eventually to marriage, its attendant financial concerns and an interest in politics. It should be noted that the full list of word frequencies broken down by age indicates that almost all words increase or decrease monotonically with age. (The few exceptions – shown in Tables 5 and 6 – involve a peak among 20s in concern with issues closely tied to student life and a sharp temporary decrease among 20s in talk about family life.) Interestingly, in almost all cases (exceptions are sports and games), words that increase monotonically with age are "male" words and those that decrease with age are "female" words.

The rapid decline in use of words expressing negative emotions confirms earlier findings [8] on a much smaller corpus.

feature	male	female
linux	**0.53±0.04**	0.03±0.01
microsoft	**0.63±0.05**	0.08±0.01
gaming	**0.25±0.02**	0.04±0.00
server	**0.76±0.05**	0.13±0.01
software	**0.99±0.05**	0.17±0.02
gb	**0.27±0.02**	0.05±0.01
programming	**0.36±0.02**	0.08±0.01
google	**0.90±0.04**	0.19±0.02
data	**0.62±0.03**	0.14±0.01
graphics	**0.27±0.02**	0.06±0.01
india	**0.62±0.04**	0.15±0.01
nations	**0.25±0.01**	0.06±0.01
democracy	**0.23±0.01**	0.06±0.01
users	**0.45±0.02**	0.11±0.01
economic	**0.26±0.01**	0.07±0.01
shopping	0.66±0.02	**1.48±0.03**
mom	2.07±0.05	**4.69±0.08**
cried	0.31±0.01	**0.72±0.02**
freaked	0.08±0.01	**0.21±0.01**
pink	0.33±0.02	**0.85±0.03**
cute	0.83±0.03	**2.32±0.04**
gosh	0.17±0.01	**0.47±0.02**
kisses	0.08±0.01	**0.28±0.01**
yummy	0.10±0.01	**0.36±0.01**
mommy	0.08±0.01	**0.31±0.02**
boyfriend	0.41±0.02	**1.73±0.04**
skirt	0.06±0.01	**0.26±0.01**
adorable	0.05±0.00	**0.23±0.01**
husband	0.28±0.01	**1.38±0.04**
hubby	0.01±0.00	**0.30±0.02**

Table 3 Word frequency (per 10000 words) and standard error by gender

feature	male	female
money	**43.6±0.4**	37.1±0.4
job	**68.1±0.6**	56.5±0.5
sports	**31.2±0.4**	20.4±0.2
tv	**21.1±0.3**	15.9±0.2
sleep	18.4±0.2	**23.5±0.2**
eating	23.9±0.3	**30.4±0.3**
sex	32.4±0.4	**43.2±0.5**
family	27.5±0.3	**40.6±0.4**
friends	20.5±0.2	**25.9±0.3**
pos-emotions	248.2±1.9	**265.1±2**
neg-emotions	159.5±1.3	**178±1.4**

Table 4 Word class frequency (per 10000 words) and standard error by gender

feature	10s	20s	30s
maths	**1.05±0.06**	0.03±0.00	0.02±0.01
homework	**1.37±0.06**	0.18±0.01	0.15±0.02
bored	**3.84±0.27**	1.11±0.14	0.47±0.04
sis	**0.74±0.04**	0.26±0.03	0.10±0.02
boring	**3.69±0.10**	1.02±0.04	0.63±0.05
awesome	**2.92±0.08**	1.28±0.04	0.57±0.04
mum	**1.25±0.06**	0.41±0.04	0.23±0.04
crappy	**0.46±0.02**	0.28±0.02	0.11±0.01
mad	**2.16±0.07**	0.80±0.03	0.53±0.04
dumb	**0.89±0.04**	0.45±0.03	0.22±0.03
semester	0.22±0.02	**0.44±0.03**	0.18±0.04
apartment	0.18±0.02	**1.23±0.05**	0.55±0.05
drunk	0.77±0.04	**0.88±0.03**	0.41±0.05
beer	0.32±0.02	**1.15±0.05**	0.70±0.05
student	0.65±0.04	**0.98±0.05**	0.61±0.06
album	0.64±0.05	**0.84±0.06**	0.56±0.08
college	1.51±0.07	**1.92±0.07**	1.31±0.09
someday	0.35±0.02	**0.40±0.02**	0.28±0.03
dating	0.31±0.02	**0.52±0.03**	0.37±0.04
bar	0.45±0.03	**1.53±0.06**	1.11±0.08
marriage	0.27±0.03	0.83±0.05	**1.41±0.13**
development	0.16±0.02	0.50±0.03	**0.82±0.10**
campaign	0.14±0.02	0.38±0.03	**0.70±0.07**
tax	0.14±0.02	0.38±0.03	**0.72±0.11**
local	0.38±0.02	1.18±0.04	**1.85±0.10**
democratic	0.13±0.02	0.29±0.02	**0.59±0.05**
son	0.51±0.03	0.92±0.05	**2.37±0.16**
systems	0.12±0.01	0.36±0.03	**0.55±0.06**
provide	0.15±0.01	0.54±0.03	**0.69±0.05**
workers	0.10±0.01	0.35±0.02	**0.46±0.04**

Table 5 Word frequency (per 10000 words) and standard error by age

	10s	20s	30s
money	31.2±0.5	44.4±0.6	**48.6±**
job	39.1±0.6	74.1±1.0	**80.6±1.9**
sports	**31.6±0.6**	24.2±0.5	23.2±0.8
tv	18.2±0.3	**19.4±0.4**	17.6±0.6
sleep	**24.1±0.4**	18.5±0.3	15.3±0.5
eating	23.6±0.4	**29.9±0.5**	27.4±0.9
sex	**42.8±0.7**	34.9±0.6	31.5±0.9
family	38.4±0.6	31.0±0.5	**41.1±1.2**
friends	**25.4±0.4**	22.7±0.4	17.0±0.4
pos-emotions	**260.7±3.0**	253.0±3.0	244.3±4.7
neg-emotions	**196.9±2.4**	152.4±1.9	139.4±2.9

Table 6 Word class frequency (per 10000 words) and standard error by age

AUTOMATED AUTHOR PROFILING

While we have found significant differences among bloggers of different ages and genders, the truest test of the significance of these differences is the extent to which they enable us to correctly predict an author's age and gender.

In this section, we will compare the extent to which author age and gender can be predicted on the basis of stylistic and content-based features, respectively. It should be noted that these classification tasks are complicated by a number of special characteristics of the blog genre. Blogs are relatively short and they often include copious quoted material (which we do not distinguish from the rest of the blog). Moreover, we must consider the possibility that some of the personal profiles provided by bloggers are bogus.

Features

For the purposes of learning age and gender classifiers, each document is represented as a numerical vector in which each entry represented the frequency (normalized by document length) of a corresponding feature in some feature set.

In earlier work, a variety of stylistic feature types have been considered for authorship attribution problems. These include function words, i.e., words that are content independent [6], syntactic features [2,10], or complexity-based feature such as word and sentence length [11]. In particular, the combination of function words with parts-of-speech has proved to be quite successful [1,5].

The stylistic features that we use are those mentioned earlier: parts-of-speech, function words, blog words and hyperlinks. There are 502 such stylistic features in all. Our content features are the 1000 unigrams with highest information gain in the training set.

We should at least note the obvious: blogs contain other clues to bloggers' age and gender, namely, blogger-provided profile information, as well as more subtle clues, such as choice of formatting template and color and use of emoticons and other non-lexical features. We do not use this information, although it is, in fact, quite useful [4].

Learning Algorithm

We use the learning algorithm Multi-Class Real Winnow (MCRW) to learn models that classify blogs according to author gender and age, respectively. Since this algorithm is not well known, we describe it briefly.

For each class, c_i, i=1,…,m, $\mathbf{w}^i$ is a weight vector $<w^i_1,\ldots,w^i_n>$, where n is the size of the feature set. Each w^i_j, is initialized to 1. Training examples are randomly ordered and processed one at a time. For a training example, $\mathbf{x} = <x_1,\ldots,x_n>$, let c_t be the actual class of $\mathbf{x}$ and let c_o (the "output" class) be $\text{argmax}_i\ \mathbf{x}*\mathbf{w}^i$. Then, if t≠o, we update as follows:

$w^t_j \leftarrow w^t_j\,(1+\beta x_j)$

$w^o_j \leftarrow w^o_j\,/(1+\beta x_j)$

where β is a positive learning constant.

We run repeated training cycles, randomly re-ordering examples after each cycle. After each ten cycles, we check the number of correctly classed training examples. If this number has diminished, we backtrack. If no improvement is found after five rounds of ten cycles each, the algorithm is terminated.

For learning problems of the scale we consider here, MCRW is much more efficient than SVM and we have found that it yields comparable results on a vast number of text categorization problems.

Gender Results

Accuracy results of 10-fold cross-validation experiments on classifying test documents according to gender are shown on the left side of Figure 1.

The results for gender using style-related features are roughly consistent with those found in earlier studies on much smaller corpora of fiction and non-fiction writing [1,5]. It is interesting to note that, despite the strong stereotypical differences in content between male and female bloggers seen above, stylistic differences remain more telling than content differences. Using all features, we obtain accuracy of 80.1%.

Age Results

Age experiments were run on the three categories considered above: 10s (13-17), 20s (23-27) and 30s (33-42). We used these categories – omitting intermediate ones – to permit clear differentiation, particularly given the fact that many of the blogs we used had been active for several years.

It should be noted that simply predicting majority class (10s) would yield accuracy of 43.8%, so this can be taken as a plausible baseline for the age experiment. Results are shown on the right side of Figure 1. The confusion matrix is shown in Table 7.

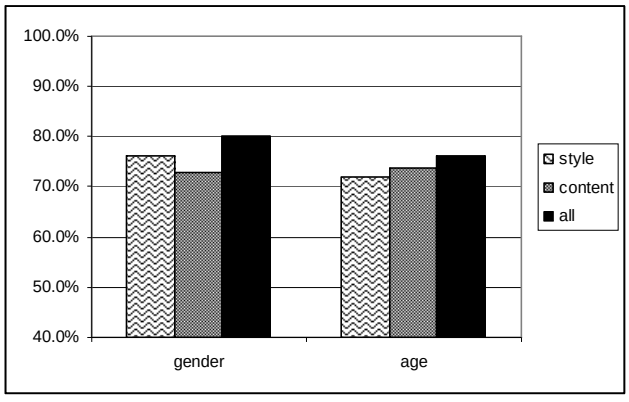

Figure 1 10-fold cross validation results for the age and gender classifiers

Classed as →	**10's**	**20's**	**30's**
10's	7036	1027	177
20's	916	6326	844
30's	178	1465	1351

Table 7 Confusion matrix for the age classifier using all features

For age, content proves to be slightly more useful than style, but – as in gender – the combination is most useful. The confusion matrix indicates that, using content and style features together, 10s are distinguishable from 30's with accuracy above 96% and distinguishing 10s from 20s is also achievable with accuracy of 87.3%. Many 30s are mis-classed as 20s, however, yielding overall accuracy is 76.2%

CONCLUSIONS

We have assembled a large corpus of blogs labeled for a variety of demographic attributes. This large sample permits us insight into the demographic distribution of bloggers. We have found that teenage bloggers are predominantly female, while older bloggers are predominantly male. Moreover, within each age group, male and female bloggers blog about different thing and use different blogging styles.

Male bloggers of all ages write more about politics, technology and money than do their female cohorts. Female bloggers discuss their personal lives – and use more personal writing style – much more than males do. Furthermore, for bloggers of each gender, a clear pattern of differences in content and style over age is apparent. Regardless of gender, writing style grows increasingly "male" with age: pronouns and assent/negation become scarcer, while prepositions and determiners become more frequent. Blog words are a clear hallmark of youth, while the use hyperlinks increases with age. Content also evolves with age in ways that could have been anticipated.

The stylistic and content differences we have found are each sufficient to permit reasonably accurate automated classification according to gender and age bracket , respectively. In each case, the combination of stylistic and content features offers the best classification accuracy.

Those who are interested in automatically profiling bloggers for commercial purposes would be well served by considering additional features – which we deliberately ignore in this study – such as author self-identification. The results reported here are relevant for cases where such features are unavailable or unreliable and for establishing a lower bound on achievable accuracy even where additional information is available. In addition, we believe that the learned models obtained in this work can be used to identify some cases in which bloggers have deliberately provided bogus profile information.

		10s	*20s*	*30s*	*all*
pronouns	*all*	**1316.7**	1173.7	1104.4	
	male	1216.4	1063.0	968.7	1113.8
	female	1416.9	1284.5	1240.1	**1334.1**
assent	*all*	**33.7**	20.1	17.0	
	male	30.0	18.5	15.3	22.9
	female	37.5	21.7	18.7	**28.0**
negation	*all*	**162.0**	157.5	149.3	
	male	153.4	146.7	137.8	148.1
	female	170.7	168.4	160.8	**168.2**
determiners	*all*	488.9	619.9	**671.5**	
	male	542.1	661.9	715.4	**619.1**
	female	435.7	578.0	627.6	525.0
prepositions	*all*	1077.0	1231.9	**1276.6**	
	male	1123.5	1250.8	1296.7	**1203.6**
	female	1030.5	1212.9	1256.5	1141.8
blogwords	*all*	**122.1**	34.8	20.4	
	male	99.2	31.3	18.7	58.3
	female	145.1	38.4	22.1	**81.4**
hyperlinks	*all*	**20.7**	35.0	38.8	
	male	25.4	41.7	49.1	35.9
	female	16.0	28.4	28.6	**23.1**
post length	*all*	195.0	210.0	**221.0**	
	male	191.4	207.5	204.1	201.0
	female	198.8	213.6	240.3	**213.0**

Table 2 Frequency (per 10,000 words) of stylistic features per gender per age bracket. (10s = 13-17; 20s = 23-27; 30s = 33-42). For all features (but post length) numbers are average of normalized per document frequency over all documents in class.

References

Argamon, S., M. Koppel , J. Fine and A. Shimoni (2003), Gender, Genre, and Writing Style in Formal Written Texts, Text , 23(3), 2003

Baayen, H., H. van Halteren, F. Tweedie. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution, Literary and Linguistic Computing, 11, 1996.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-linguistic Comparison* (Cambridge University Press, Cambridge)

de Vel O., M. Corney, A. Anderson and G.Mohay (2002). "Language and Gender Author Cohort Analysis of E-mail for Computer Forensics", Digital Forensic Research Workshop, August 7 – 9, 2002, Syracuse, NY

Koppel, M., S. Argamon, A. Shimony (2002). Automatically categorizing written texts by author gender, Literary and Linguistic Computing 17,4, (2002), 401-412

Mosteller, F. and Wallace, D. L. (1964). Inference and Disputed Authorship: The Federalist. Reading, Mass. : Addison Wesley, 1964.

Mulac, A., J.J. Bradac and P. Gibbons (2001). Empirical support for the gender-as-culture hypothesis: An intercultural analysis of male/female language differences, Human Communication Research 27:121-152 (2001)

Pennebaker, J.W., & Stone, L.D. (2003). Words of wisdom: Language use over the lifespan. Journal of Personality and Social Psychology, 85, 291-301 (2003).

Pennebaker, J.W., Francis, M.E., and Booth, R.J. (2001). Linguistic Inquiry and Word Count: LIWC 2001. Mahwah, NJ: Erlbaum Publishers

Stamatatos, E., N. Fakotakis & G. Kokkinakis (2001). Computer-based authorship attribution without lexical measures, Computers and the Humanities 35, (2001) 193—214.

Yule, G.U. (1938). On Sentence Length as a Statistical Characteristic of Style in Prose with Application to Two Cases of Disputed Authorship, Biometrika, 30, (1938) 363-390.