Stylometric Analysis of Bloggers' Age and Gender

Sumit Goswami **Sudeshna Sarkar** **Mayur Rustagi**

Department of Computer Science and Engineering, Indian Institute of Technology (IIT) Kharagpur
West Bengal 721 302, India
sgoswami@iitkgp.ac.in, sudeshna@cse.iitkgp.ernet.in, mrustagi@iitkgp.ac.in

Abstract

We report results of stylometric differences in blogging for gender and age group variation. The results are based on two mutually independent features. The first feature is the use of slang words which is a new concept proposed by us for Stylometric study of bloggers. Slang is a non-dictionary word that has evolved with time due to its frequent and popular usage. For the second feature, we have analysed the variation in average length of sentences across various age groups and gender. These two features are then augmented with previous study results reported in literature for stylometric analysis of age and gender. The combined feature list enhances the accuracy by a remarkable extent in predicting age and gender. These experiments were done on a 20,000 blog corpus. Experimental results show that these features work well in detection of bloggers' demography. However, gender determination is more accurate than age group detection over a data spread across all ages but the accuracy of age prediction increases if we sample data with remarkable age difference.

Introduction

Gender and age are the common demographic features used for experimentation using stylometry as the blogs generally contain these information provided by the author. Style in writing is a result of the subconscious habit of the writer of using one form over a number of available options to present the same thing. The variation also evolves with the usage of the language in certain period, genre, situation or individuals. Variation are of two types – variation within a norm which is grammatically correct and deviation from the norm which in ungrammatical. The variations can be described in linguistic as well as statistical terms (McMenamin 2002). Concept and themes (Leximancer 2008; Weber 1990) can be determined from variations within the norm while usage of non-dictionary words or slang is an example of deviation from a norm.

Blogs substantially reduced the technical and language skills required to publish. It has brought forward a wide variety of reporting techniques, content type, style and goals of blogging. Bloggers generally express their thoughts in an informal, unreserved and unorganized manner through the blogs. The language used here has a mixed characteristic of spoken and written language constructs like use of jargons, abbreviations, too many exclamations, short sentences, emotion symbols etc. The topics which were considered private are openly discussed by the teenagers and young adults (Mishne 2006).

Related Work

The research in last few decades on usage of language pattern by different social groups was constrained due to unavailability of sufficient and annotated data. The growth of blogosphere with its availability for electronic downloading has simplified the data collection. Analyses of effects of bloggers' age and gender from weblogs have been presented by a few (Schler et al. 2006; Burger and Henderson 2006; Yan 2006 ; Argamon, Koppel and Avneri 1998; Yan and Yan 2006; Nowson and Oberlander 2006) but as per our analysis these are generally based on usage of keywords, parts of speech and other grammatical constructs. More work has been done of gender estimation than age determination. Age linked variations had been reported by (Pennebaker et al. 2001; (Pennebaker and Stone 2003; Burger and Henderson 2006). (Koppel, Argamon and Shimoni 2003) estimated author's gender using the British National Corpus text. By using function words and part-of-speech, (Schler et al. 2006) reported 80% accuracy for classifying author's gender. It also stated that female authors tend to use pronoun with high frequency, and male authors tend to use numeral and representation related numbers with high frequency.

Data

A blog corpus is available on the website of Prof Moshe Koppel (Schler et al. 2006). It has collection of blogs from blogger.com collected in August 2004. It reports to have collected all 71493 accessible blogs on the site which (a) contained at least 500 total words including at least 200 occurrences of common English words, and (b) had author–provided indication of both gender and age. It has 681288 blog posts from 19320 bloggers written from January 1999 till the date of data collection. From this collection, 9660 of male and female blogs each were filtered out. However, in this corpus, teenage blogs greatly outnumber the adult blogs. For our experiments we had

used this corpus either completely or selectively as per experimental requirements.

Features

The highest frequency words collected from a corpus may not be a good distinguishing feature. But an analysis of the words that are highest occurring in a sub-corpus can be the marker (Datta and Sarkar 2008). Reference to 'attending school' results in an instant 'teenage' classification. A feature may be represented by its relative frequency or by its mere presence or absence. Features for stylometrics are generally based on character or morphological features or lexical features. In our experiments we used the sentence length and non-dictionary words as the features. As per our literature survey, the usage of slang word has not yet been explored for study of stylometric variation.

Sentence Length

Figure 1 shows the variation of average sentence length on age and gender basis. The age bracket of 10s, 20s and 30s represent the age group of 13-17, 23-27 and 33-42 respectively. We selected to work on this feature because we found, most of the reported work was on formal writing and generally on classical works of literature. Analysis of blogs based on average sentence length is challenging as blogs lacks editorial and grammatical checks.

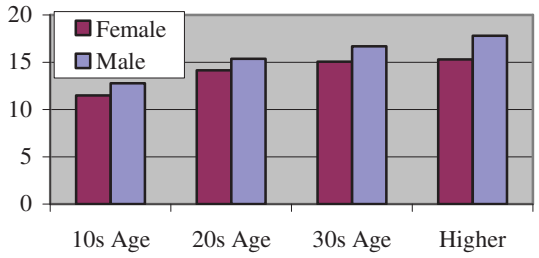

Figure 1 Average sentence length on Gender and Age basis

Non-Dictionary Words

As blogs are informal writing without any editorial bounds, it has slowly filled up with many non-dictionary words that are understandable and commonly used by online community. We refer to some of them as slangs, smiley, out-of-dictionary words, chat abbreviations etc. The named entities are also non-dictionary words. There are words that are intentionally misspelled, repeated, extended or shortened to have a different effect on the reader, express emotion or save the time of blogging. All these words and the frequency of use of such words are contributable features in stylometrics. Figure 2 shows the usage of non-dictionary words among age and gender variation and Table 1 shows the usage frequency of a few selected non-dictionary words among different gender and age groups respectively

Results and Discussion

Non-Dictionary Words

Analysis of Figure 2 tells that teenagers generally use more non-dictionary words than the adults. Here, we call those words as non-dictionary which is not available in the Ispell ver 3.1.20. Though, the number of slang words used in text can be a remarkable feature but a single feature can't make a good classifier. To build a classifier for age variation, we initially took only those bloggers who are in their 10s and those who are in their 30s so that there is a remarkable difference between their usage of non-dictionary word pattern and thus simpler to classify.

For our experiments with non-dictionary words, only those words were selected as feature which had an occurrence of >50 and for which the usage among male and female was atleast double. 52 words were found and used, a partial list of which is given in Table 1.

Naïve Bayes Classifier yielded an accuracy of 77.39 % for gender based classification and 89.68 % accuracy for the age group classification between 10s and 30s age. The confusion matrix of gender and age linked classification is given in Table 2 and Table 3 respectively

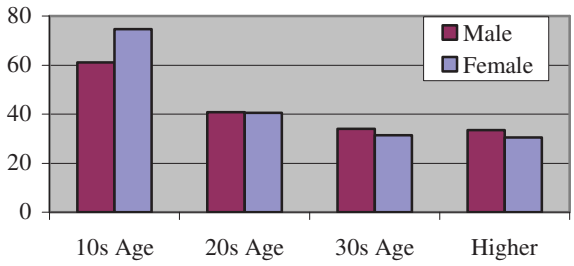

Figure 2 Non-Dictionary words used per 1000 words across various age groups

Table 1 : Partial List of non-dictionary word frequency per 100000 words in gender and age class

Non-dictionary Word	Gender Variation		Age Variation	
	Male	Female	10s Age	30s Age
Yay	10.7294	44.3713	35.5822	17.1137
wat	12.433	57.5201	4.5026	20.7462
tv	21.6521	56.5894	37.3394	29.7338
Thats	10.4031	20.4301	207.2521	27.2996
thats	41.2864	176.967	207.2521	27.2996
sunday	14.6321	42.0974	30.1094	9.2871
saturday	14.0642	44.6894	32.2876	8.5007
sa	20.5888	76.5011	37.5407	9.1373
ok	74.8883	295.5773	252.0228	97.3277
nite	11.7564	41.143	37.0831	8.7628
ng	11.5147	58.0621	31.7568	8.0513
na	30.3999	117.2907	73.123	12.4327

mins	10.681	24.2711	18.9076	9.3245
hmmm	13.5204	28.1121	33.1112	11.6463
hehe	25.3494	92.9843	97.1008	7.8641
fuckin	11.7926	23.7409	23.8679	10.3356
didnt	23.8149	144.5309	165.4466	21.3079
da	12.6142	63.1167	70.7069	21.5326
Cuz	18.2327	195.6416	208.9726	28.7601
A lot	28.1646	85.5615	84.2699	30.5576
Lol	18.5348	331.5598	439.4521	47.4467
I've	58.6733	168.4485	104.8066	31.8309
Ish	10.246	38.3742	31.3175	14.8669
Im	20.8063	85.6675	92.0307	21.8697
I'm	165.2545	594.6422	5.6924	23.9293
Im	46.8323	520.6506	661.4935	18.3495
I'll	42.4463	152.0361	132.3353	21.5326
I'd	20.613	60.1829	35.1795	10.1484
Hmm	17.0607	49.1902	36.3876	17.2635
hmm	9.944	51.2638	54.4716	8.3509
friday	17.2057	59.0636	46.5827	10.2233
everytime	10.3669	31.7998	24.6183	9.9237
english	11.3577	36.4773	36.5157	7.7143
dont	64.8597	372.8679	436.011	56.4343
doesnt	10.5844	53.6674	58.9194	10.785

Table 2 Confusion matrix for the gender classification using 52 non-dictionary words as features

a	**b**	← classified as
4916	431	**a = male**
1988	3366	**b = female**

Table 3 Confusion matrix for 10s and 30s age group classification using 52 non-dictionary words as features

a	**b**	← classified as
7136	**1104**	**a = 10s age**
0	**2461**	**b = 30s age**

Average Sentence Length

Koppel in his paper (Schler et al. 2006) used a list of 30 words each as a distinguishing feature for gender and age respectively. These words, which we refer here as 'content words', were detected to be having an extreme variation in usage across gender and age groups. Though, average sentence length is a remarkable feature but a single feature can't make a good classifier. So we used this feature in combination with slang words reported above and the content words.

As blogs are informal writing, the bloggers' may not abide by the grammatical and editorial rules and can use huge sentences or end up in a 2-3 word sentence to just impart the meaning. There are grammatical errors in the blog writing like improper use of full stop (.), exclamation marks (!) or capital letters.

The classification results and Figure 1 is not sufficient to interpret that the average sentence length in a persons writing increases with age. The blogposts collected in the corpus had been written across a span of about five years, which is not sufficient to predict this trend. The trend of increase in the average sentence length with age can be tested only if we have sufficient blog data in which the person had been blogging for a few decades so as to look into the trend of change in average sentence length with his age. It may happen that the average sentence length in English writing is decreasing with time. Those who are blogging today may continue blogging at the same average sentence length but those who start blogging after ten years may use further smaller sentence lengths.

Augmented Features

Age experiments were run on four categories of age group: 10s, 20s, 30s and higher. The feature list comprised of 35 content words combined with 52 slang words mined by us from blog data based on our acceptance index. (Schler et al. 2006) has reported an accuracy of 76.2% with the content words. The augmented feature list yielded an accuracy of 80.32%. The confusion matrix is given in Table 4. Addition of average sentence length to this set of features further increased the accuracy by a small amount to 80.38 %. The confusion matrix and the detailed accuracy by class for are given in Table 5 and Figure 3.

Similarly, experiment was done for gender variation after augmenting the 35 content words with 52 slang words. (Schler et al. 2006) has reported an accuracy of 80.1 % in gender determination. Our augmented feature list gave an accuracy of 89.18%, the confusion matrix of which is given in Table 6. After augmenting feature list with average sentence length, there was an increase in the accuracy to 89.30 %. The confusion matrix and the accuracy by class are given in Table 7 and Figure 4. With these results it should not be interpreted looking at Figure 1 that the average sentence length increases with age.

Table 4: Confusion matrix for the age classifier using 52 slang words and 35 content words

a	**b**	**c**	←classified as
7334	232	674	**a = 10**
0	5327	2759	**b = 20**
0	31	2430	**c = 30**

Table 5: Confusion matrix for age classifier using 52 slangs, 35 content words and average sentence length

a	**b**	**c**	←classified as
7342	230	668	**a = 10**
5	5330	2751	**b = 20**
0	31	2429	**c = 30**

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.891	0.001	0.999	0.891	0.942	0.991	10s
0.659	0.024	0.953	0.659	0.779	0.948	20s
0.987	0.209	0.415	0.987	0.585	0.914	30s

Figure 3 Accuracy By class for age group detection

Table 6 Confusion matrix for the gender classifier using 52 slangs and 35 content words

A	**b**	← classified as
9660	0	**a = male**
2089	7571	**b = female**

Table 7: Confusion matrix for gender classifier using 52 slangs, 35 content words & average sentence length

a	**b**	← classified as
9660	0	**a = male**
2067	7593	**b = female**

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.214	0.824	1	0.903	0.98	male
0.786	0	1	0.786	0.88	0.98	female

Figure 4 Detailed Accuracy By Class gender detection

Conclusion and Future Work

Teenage bloggers use more out-of-dictionary words than the adult bloggers. Furthermore, for bloggers of each gender, there is a clear distinction between usages of a few slangs. In their present age, teenager use smaller sentences compared to the adult bloggers. With the available data and the existing experiments, it cannot be confirmed that the average sentence length increases with age.

The stylistic difference in usage of slang predicts the age and gender variation with certain accuracy. Average sentence length in itself is not a good feature to predict the variation as there is a wide variation in sentence length in informal writing. However, the feature of average sentence length can be augmented with slangs to slightly increase its prediction efficiency. Both these features when augmented with other features like content words reported earlier, increases the prediction accuracy by a good amount.

The usage of slang can be a good feature to predict the geographical location or the ethnic group of the user. We also require a sufficiently huge corpus collected over a span of more than ten years to determine the variation of sentence length with age. This can be used to study individuals' language use and changes in it over the course of their lives. This corpus can also be used to study the evolution and death of the slang words with time.

References

Argamon, S., Koppel, M. and Avneri, G. 1998 *Routing Documents According to Style* First International Workshop on Innovative Information Systems, 1998.

Burger, J. D., and Henderson, J. C. 2006 *An Exploration of Observable Features Related to Blogger Age*. Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs, 2006.

Datta, S., Sarkar, S. 2008 *A Comparative Study of Statistical Features of Language in Blogs-vs-Splogs*, Proceedings of the second workshop on Analytics for noisy unstructured text data, Singapore, 63-66.

Koppel, M.; Argamon, S.; and Shimoni, A. R. 2003. *Automatically categorizing written texts by author gender*. Literary and Linguistic Computing, 14, 2001

Leximancer 2008, *Leximancer Manual,* Ver 3, www.leximancer.com, last accessed on January 22, 2009
McMenamin, G. R. 2002 *Forensic Linguistics : Advances in Forensic Stylistic*.: CRC Press

Mishne, G. 2006 *Information Access Challenges in the Blogspace*, IIIA-2006 – International Workshop on Intelligent Information Access, 2006

Nowson, S., and Oberlander, J. 2006. *The identity of bloggers: Openness and gender in personal weblogs*. AAAI Spring Symposium Series Computational Approaches to Analyzing Weblogs, 163–167.

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

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

Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. 2006 *Effects of Age and Gender on Blogging*. Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs, 2006.

Weber, R.P. 1990 *Basic Content Analysis*. Newbury Park, Calif.: Sage Publications, 2 ed

Yan, R. 2006 *Gender Classification of Weblog Authors with Bayesian Analysis*. Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs, 2006.

Yan, X., and Yan, L. 2006. *Gender classification of weblog authors*. AAAI Spring Symposium Series Computational Approaches to Analyzing Weblogs, 228–230