Learning Age and Gender of Blogger from Stylistic Variation

Mayur Rustagi, R. Rajendra Prasath*, Sumit Goswami, and Sudeshna Sarkar

Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur,
West Bengal 721302, India
mrustagi@iitkgp.ac.in, rajendra@cse.iitkgp.ernet.in,
sgoswami@iitkgp.ac.in, sudeshna@cse.iitkgp.ernet.in

Abstract. We report results of stylistic 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 Stylistic study of bloggers. For the second feature, we have analyzed the variation in average length of sentences across various age groups and gender. These features are augmented with previous study results reported in literature for stylistic analysis. The combined feature list enhances the accuracy by a remarkable extent in predicting age and gender. These machine learning experiments were done on two separate demographically tagged blog corpus. Gender determination is more accurate than age group detection over the data spread across all ages but the accuracy of age prediction increases if we sample data with remarkable age difference.

1 Introduction

Stylistic classification can improve the results achieved through Information Retrieval (IR) techniques by identifying documents that matches a certain demographic profile. Gender and age are the common demographic features used for experimentation using stylistics 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. Variations are of two types - variation within a norm which is grammatically correct and deviation from the norm which is ungrammatical. The variations can be described in linguistic as well as statistical terms[15]. Concept and themes[20] can be determined from variations within the norm while the usage of non-dictionary words or *slang* is an example of deviation from a norm.

2 Related Work

The research in last few decades on usage of language pattern by different social groups was constrained due to unavailability of sufficient annotated data.

* Currently Rajendra is an ERCIM Fellow at IDI, NTNU, Norway.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 205–212, 2009.

Analysis of effects of bloggers age and gender from weblogs, based on usage of keywords, parts of speech and other grammatical constructs, has been presented in [2,6,19,22]. Age linked variations had been reported by Pennebaker, et al. [11], Pennebaker and Stone[14] and Burger and Henderson, 2006 [6]. J. Holmes distinguished characteristics of male and female linguistic styles [9]. Expert used spoken language [15], Palander worked on electronic communications[13], and S. Herring analyzed correspondence[18]. Simkins analysed that there are no difference between male and female writing style in formal contexts [10]. Koppel et al. estimated author's gender using the British National Corpus text [12]. By using function words and part-of-speech, Koppel et al. reported 80% accuracy for classifying author's gender. Koppel et al. 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. Corney et al. estimated author's gender from e-mail content [5]. In addition to function words and part-of-speech and n-grams [19,12], they used HTML tags, the number of empty lines, average length of sentences for features for SVM [4].

3 Dataset

A blog corpus[1] is made available by ICWSM 2009[1] and the blogs in this corpus did not have any tag for demographic information. However, it had the resource link which had the URL of the blogger's home page. In the above corpus, blogs from blog.myspace.com had the maximum occurrence and had the demographic details of the blogger in its home page. The home page of these URLs were crawled and processed to retrieve gender, status (married, unmarried), age, zodiac sign, city, state and country corresponding to each URL. With the available valid URL list, the downloaded data from these URLs gives 92,381 files. The blogs in which the blogger's age has been reported as below 20

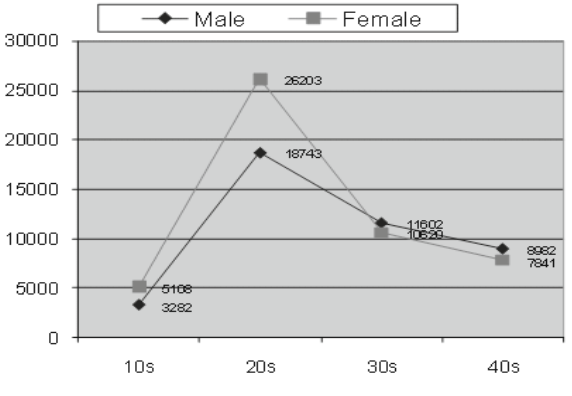

Fig. 1. Number of files in age groups and gender

[1] Provided by Spinn3r.com [3].

has been grouped in 10s age group, those in the age group of 20 to 29 as 20s, those in 30 to 39 as 30s and 40 and above has been put in 40s age group. The distribution of these files over age and gender is given in Figure 1.

4 Feature Selection and Classification Algorithm

The highest frequency words collected from a corpora may not be a good distinguishing feature. However, an analysis of the words that are highest occurring in a subcorpora can be the marker [16]. 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 stylistics 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 the study of stylistic variation.

Koppel[12] used a list of 30 words each as a distinguishing feature for gender and age respectively. These words were detected to be having an extreme variation in usage across gender and age groups. Similarly out-of-dictionary words were augmented to increase the accuracy of results[17]. For the purpose of learning age and gender classifier, each document is represented as a numerical vector in which each entry represent the normalized frequency of a corresponding word in the feature set. Table 1 and Table 2 lists a few content words used for learning gender and age groups respectively.

Many Stylistic results had been reported using average sentence length as a feature. Still we selected to work on this feature because, 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 lack editorial and grammatical checks. Figure 2 shows the variation of average sentence length on age and gender basis.

As blogs are informal writing without any editorial bounds, blogosphere 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 vocabulary 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 even the frequency of use of such words are contributable features in stylistics. A taboo word for a particular age group can be a commonly used word for another. For our experiments with non-dictionary words, Ispell [8] was run and the frequency of all the non-dictionary words used by males and females for detecting gender variation was obtained. From these, 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. This generated a list of 52 words. Figure 3 shows the usage of out of vocabulary words among age and gender variation and Figure 4 shows the usage frequency of a few selected Non-Dictionary words among different gender.

Table 1. List of Content word frequency per 10000 words in gender

	Male Occ / 10000	Female Occ / 10000
mom	4.543	7.844
microsoft	0.921	0.594
gaming	0.131	0.045
server	0.152	0.108
software	0.131	0.051
gb	0.436	0.519
programming	0.069	0.045
google	0.318	0.228
data	0.249	0.114
graphics	0.076	0.108
india	0.069	0.085
nations	0.464	0.142
democracy	0.048	0.011
users	0.159	0.102
economic	0.159	0.079
shopping	0.304	0.845
cried	0.159	0.759
freaked	0.048	0.119
pink	0.256	0.497
cute	0.671	1.662
gosh	0.083	0.182
kisses	0.096	0.217
yummy	0.069	0.091
mommy	0.027	0.314
boyfriend	0.297	1.411
skirt	0.062	0.154
adorable	0.027	0.285
husband	0.297	1.765
hubby	0.034	0.359

Table 2. List of Content word frequency per 10000 words in age groups

	$\frac{WC}{(\sum \text{ WC in that age grp})\text{x}10000}$		
Word	10s age	20s age	30s age
college	4.433	1.173	0.829
maths	0	0.006	0
homework	0.299	0.126	0.078
bored	2.399	1.892	0.789
sis	3.433	4.750	4.844
boring	0.966	0.687	0.618
awesome	2.533	2.971	2.264
mum	0.499	0.277	0.329
crappy	0.266	0.283	0.289
mad	9.832	9.236	8.384
dumb	1.266	0.870	0.447
semester	1.333	0.813	0.263
apartment	0.599	1.205	0.487
drunk	0.799	1.318	0.974
beer	0.466	0.826	0.908
student	0.766	0.504	0.855
album	0.966	1.463	1.684
someday	0.199	0.302	0.184
dating	0.699	0.889	0.710
bar	3.733	3.470	3.922
marriage	0.133	0.403	0.394
development	0.099	0.176	0.171
campaign	0.033	0.258	0.605
tax	0.066	0.391	0.539
local	0.499	0.706	1.803
democratic	0.033	0.044	80.10
son	30.26	28.80	28.55
systems	0	0.050	0.105
provide	0.433	0.378	0.552
workers	0.099	0.233	0.394

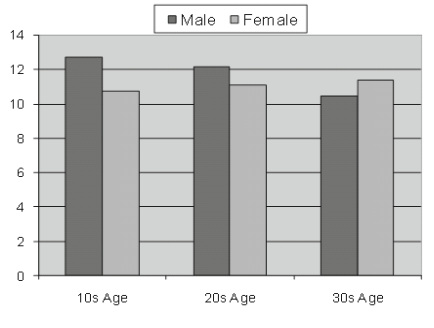

Fig. 2. Average Sentence length on Gender Age Basis

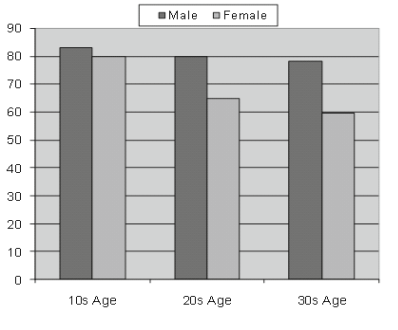

Fig. 3. Out-of-dictionary words used per 1000 words across various age groups

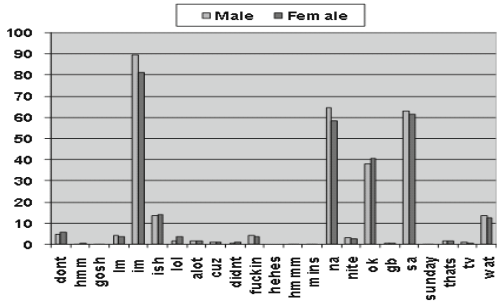

Fig. 4. Measure of usage of a few selected Non-dictionary words used by Males and Females

Naive Bayes classifier for predicting the blogger's age group or gender from the stylistic features were trained using the WEKA toolkit [21]. During training, classifiers are created by the selection of a set of variables for each feature and classifier parameters are tuned through cross-validation. To evaluate the classifier, the given data is split into training and test data and the trained classifier is used to predict the blogger's demographic profile on the test data [7].

5 Results and Discussion

Analysis of Figure 3 tells that teenagers generally use more out-of-dictionary words than the adults. Here, we call those words as non-dictionary which are not available in the Ispell ver. 3.1.20. These words can be the slang, exclamatory word with a small or large repetition of last character like 'soooo sweet' or typing errors due to less concern towards spelling and grammar or idiosyncrasies. Though, the number of slang words used in text can be a remarkable feature but a single feature can not 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, we selected the list of 52 non-dictionary words.

Table 3. Confusion matrix for the gender classification using 52 non-dictionary words as features

a	b	← classified as
42609	0	a = male
15147	34608	b = female

Naive Bayes Classifier yielded an accuracy of 83.6% for gender based classification and 95% accuracy for the age group classification between 10s and 30s age. We did not measure the percent accuracy for age group classification between 10s, 20s and 30s due to similarity of style in overlapping age groups.

5.1 Average Sentence Length

Since the average sentence length is a remarkable feature, we used this feature in combination with slang words reported above and the interesting content words reported on this corpus in [19]. The classification results and Figure 2 are not sufficient to interpret that the average sentence length in a persons writing increases with age. The collected blogposts had been written across a span of 5 years and 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. The bloggers of today may continue blogging at the same average sentence length but those who start blogging after ten years may use smaller sentence lengths.

Table 4. Detailed Accuracy By Class gender detection using out-of-dictionary words only

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1.000	0.304	0.738	1.000	0.849	0.925	male
0.696	0.000	1.000	0.696	0.820	0.925	female

Table 5. Confusion matrix for 10s and 30s age group classification using 52 non-dictionary words

a	b	← classified as
6890	1498	a = 10s age
3	22219	b = 30s age

5.2 Augmented Features

The gender and age experiments were conducted initially only on 35 content words and it gave an accuracy of 95.2% for gender classification, and 92.51% for age classification (refer to Table 3).

The age experiments were run on four categories of age group considered above: 10s, 20s, 30s and higher. The feature list comprised of 35 content words reported in [19] combined with 52 slang words mined by us from blog data based on our acceptance index. [19] has reported an accuracy of 92.51% with the content words. The augmented feature list yielded an accuracy of 94.13%.

Table 6. Confusion matrix for the gender classification using 35 Content words as features

a	b	← classified as
42571	0	a = male
4451	45262	b = female

Table 7. Detailed Accuracy By Class gender detection using content words only

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.09	0.905	1	0.95	1	male
0.91	0	1	0.91	0.953	1	female

Table 8. Confusion matrix for the gender classifier using 52 slang word and 35 content words

a	b	← classified as
30931	0	a = male
1731	34191	b = female

Addition of average sentence length to this set of features did not contribute to a significant amount.

Similarly, experimentation was done for gender variation after augmenting the 35 content word feature reported in [19] with our 52 slang words. Schler *et* al.[19] have reported an accuracy of 80.1% in gender determination on their dataset and we received an accuracy of 83.6% on ICWSM dataset. However, our augmented feature list gave an accuracy of 97.41%, the confusion matrix of which is given in Table 8. After further enhancement of this augmented feature list with average sentence length, there was not much increase in the accuracy and so is not reported here.

6 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 usage of a few slangs. Generally in their present age, teenager use smaller sentences compared to the adult bloggers but we found a variation to this in this dataset. With the available data and the existing experiments, it cannot be confirmed that the average sentence length increases or decreases with age. The stylistic difference in usage of slang predicts the age and gender variation with certain accuracy. Average sentence length 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 slang 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 also be a good feature to predict the geographical location or the ethnic group of the user due to the heavy usage of a particular out-of-dictionary word or named entities at certain regions. A sufficiently huge corpus collected over a decade will be useful to study the variation of sentence length of users with age and variations in individuals language use 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

1. ICWSM 2009, Spinn3r Dataset (May 2009)
2. Argamon, S., Koppel, M., Avneri, G.: Routing documents according to style. In: Proc. of First Int. Workshop on Innovative Inform. Syst. (1998)
3. Spinn3r Indexing Blogosphere, www.spinn3r.com (last accessed on March 01, 2009)
4. Brank, J., Grobelnik, M., Milic-Frayling, N., Mladenic, D.: Feature selection using support vector machines. In: Proc. of the 3rd Int. Conf. on Data Mining Methods and Databases for Eng., Finance, and Other Fields, pp. 84–89 (2002)
5. Corney, M., de Vel, O., Anderson, A., Mohay, G.: Gender-preferential text mining of e-mail discourse. In: 18th Annual Computer Security Appln. Conference (2002)
6. Burger, J.D., Henderson, J.C.: An exploration of observable features related to blogger age. In: Proc. of the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs (2006)
7. Estival, D., Gaustad, T., Pham, S.B., Radford, W., Hutchinson, B.: Tat: an author profiling tool with application to arabic emails. In: Proc. of the Australasian Language Technology Workshop, pp. 21–30 (2007)
8. Ispell (2009), http://www.gnu.org/software/ispell/ (last accessed on April 02, 2009)
9. Holmes, J.: Women's talk: The question of sociolinguistic universals. Australian Journal of Communications 20(3) (1993)
10. Simkins-Bullock, J., Wildman, B.: An investigation into relationship between gender and language Sex Roles 24. Springer, Netherlands (1991)
11. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic Inquiry and Word Count. In: LIWC 2001 (2001)
12. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
13. Palander-Collin, M.: Male and female styles in 17th century correspondence: I think. Language Variation and Change 11, 123–141 (1999)
14. Pennebaker, J.W., Stone, L.D.: Words of wisdom: Language use over the lifespan. Journal of Personality and Social Psychology 85, 291–301 (2003)
15. McMenamin, G.R.: Forensic Linguistics: Advances in Forensic Stylistic. CRC Press, Boca Raton (2002)
16. Datta, S., Sarkar, S.: A comparative study of statistical features of language in blogs-vs-splogs. In: AND 2008: Proc. of the second workshop on Analytics for noisy unstructured text data, pp. 63–66. ACM, New York (2008)
17. Goswami, S., Sarkar, S., Rustagi, M.: Stylometric analysis of bloggers' age and gender. To appear in: Proc. of ICWSM (2009)
18. Herring, S.: Two variants of an electronic message schema. In: Herring, S. (ed.) Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives, vol. 11, pp. 81–106 (1996)
19. Argamon, S., Schler, J., Koppel, M., Pennebaker, J.: Effects of age and gender on blogging. In: Proc. of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs (April 2006)
20. Leximancer Manual V.3, www.leximancer.com (last accessed on January 22, 2009)
21. Witten, I.H., Frank, E.: DataMining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
22. Yan, R.: Gender classification of weblog authors with bayesian analysis. In: Proc. of the AAAI Spring Symp. on Computational Approaches to Analyzing Weblogs (2006)