Automatic Estimation of Bloggers' Gender

Daisuke Kobayashi
The University of Tokyo
Hongo 7-3-1, Bunkyo-ku
Tokyo 113-8656 Japan
d-koba@mi.ci.i.u-tokyo.ac.jp

Naohiro Matsumura
Osaka University
Machikaneyama 1-7,
Toyonaka
Osaka, 560-0043 Japan
matumura@econ.osaka-u.ac.jp

Mitsuru Ishizuka
The University of Tokyo
Hongo 7-3-1, Bunkyo-ku
Tokyo 113-8656 Japan
ishizuka@i.u-tokyo.ac.jp

Abstract

We propose an approach employing Support Vector Machine (SVM) to estimate bloggers' gender from blog posts. The data we analyze consists of blog posts on Doblog (Japanese blog-hosting service) and questionnaire results by Doblog users. Experimental evaluations show that our approach achieved 90% accuracy for 83% bloggers.

Keywords

Gender Estimation, Text Classification, User Profiling

1. Introduction

As blogs are open to the public, people can access to any blogs to see what other bloggers are thinking of. By aggregating others' blogs, we can see the trend emerging on blogosphere in near real time. The trend is valuable for business people in establishig marketing maneuver. For example, business people in charge of product development are eager to catch the trend to produce new product reflecting bloggers' need. Also, business people in charge of marketing use the trend to propose how to place banner ads or affiliate ads on blogs and portal sites. However, such blog based approaches face a serious problem: trend for men and for women are definitely different in most domains, and the same can be said for bloggers' age, residential area etc. The problem is derived from the lack of bloggers' personal information because such information is not opened to the public in general. In this paper, we propose an approaches employing Support Vector Machine (SVM) to estimate bloggers' gender from blog posts.

2. Method

Estimating all blog posts into either male or female is not practical because not all of them have dominant features for gender estimation. To overcome this obstacle, we define a "neutral" class in addition to "male" and "female" classes, and filter out the blog posts estimated as neutral. Hereafter we call a blog post estimated as neutral as a *foggy post*.

To identify foggy posts, we propose *M-Score* and *F-Score* each of which showing the plausibility of the gender of the blogger as male or female respectively. M-Score and F-Score is based on the sum of weights of features for each feature. We employ J. Brank's approach [1] for extracting weights of features.

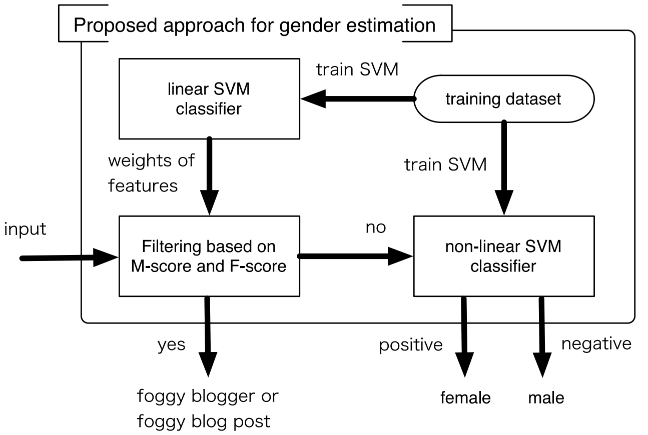
Proposed approach for gender estimation
linear SVM classifier
train SVM
training dataset
weights of features
train SVM
input
Filtering based on M-score and F-score
no
non-linear SVM classifier
yes
positive
negative
foggy blogger or foggy blog post
female
male

Fig. 1: *Overview of gender estimation process*

% Once we obtained M-Score (s_m^b) and F-Score (s_f^b), foggy posts are determined if below equation is satisfied.

$$|\log\left(s_f^b/s_m^b\right)| \le c_n \quad \text{or} \quad s_f^b \ge c_s \quad \text{or} \quad s_m^b \le c_s \tag{1}$$

c_n and c_s are the parameters for threshold.

Gender estimaion process goes as follows. First, we train linear SVM and non-linear SVM with training dataset respectively. Then, weights of all features for each blog post in test dataset are extracted to measure M-Score and F-score, and based on which foggy blogs are to be filtered out. After filtering out foggy blogs, the rest of blogs are classified by non-linear SVM to estimate authors' genders. The filtered blog posts are classified as 'neutral' class. The overview of gender estimation process is shown at Figure 1.

3. Experiment

3.1 Training and test dataset

The training and test dataset for SVM is prepared from blog posts provided by Doblog. We generate a feature vector for each blog post by arranging N-gram (one- to ten-grams) of noun, verb, and adjective words as features. Those words are selected because they can be constituent parts independently. The initial value of each feature is set by $tfidf$, and each feature vector is normalized to unit length. For each blog post, we can tell the gender exactly from questionnaire results provided by Doblog. By integrating feature vectors with questionnaire results, we can prepare supervised training/test dataset. In preparing training and test dataset, 1,000 posts

ICWSM '2007 Boulder, CO USA

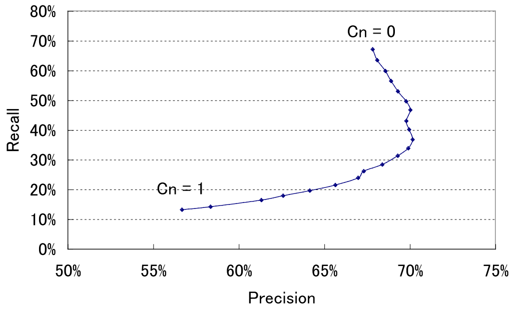

Fig. 2: *Precision-recall curve with parameter c_n*

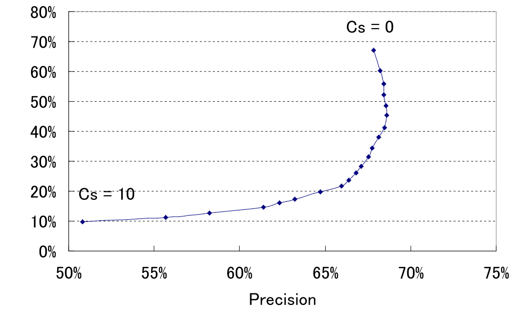

Fig. 3: *Precision-recall curve with parameter c_s*

by male and 1,000 posts by female are randomly sampled for training dataset (total 2,000 posts), and another 10,000 posts by male and 10,000 posts by male are chosen for test dataset (total 20,000). Note that the training and test dataset are obtained from disjoint bloggers.

3.2 Precision and recall for parameters

We change filtering parameters, c_n and c_s, to see precision-recall curve. Figure 2 shows precision-recall curve with c_n. If c_n gets increased from 0 to about 0.5, the precision is improved although the recall is worsened. For c_n over about 0.5, both the precision and recall are worsened. Figure 3 shows precision-recall curve with c_s. The recall is rapidly decreasing as c_s increases. The precision is stable if c_s is less than about 5, however it is rapidly decreasing for c_s over about 5.

3.3 Grouping blog posts

The task we tried so far is to estimate the author's gender for each blog post. However the task is not easy for us because of the lack of clues in a blog post in many cases. In this section, we see what happens if we first group the posts corresponding to each blogger, and consider them as *a big post*, then do the filtering, and then estimate the gender. We investigate the result of accuracy for the number of blog posts in a bundle. The accuracy increases until the number of blog posts reaches around 30, and then becomes stable for more blog posts. From the result, we conclude that it is preferable to group at least 30 blog posts when estimating the author's gender.

Filter	Param.	Accuracy	Male/Female/Total	Coverage
No filtering	-	0.811	480/272/752	1.000
DS + small	150	0.854	281/186/467	0.621
DS + frac	0.07	0.850	398/177/575	0.764
DS + margin	7	0.845	409/197/606	0.801
TS + small	3.5	0.844	390/240/630	0.838
TS + frac	0.04	0.854	438/247/685	0.911
TS + margin	0.12	0.851	441/251/692	0.920
DS + small	200	0.894	136/101/237	0.315
DS + frac	0.15	0.898	227/53/280	0.372
DS + margin	20	0.895	309/100/409	0.544
TS + small	4.2	0.899	226/169/395	0.525
TS + frac	0.08	0.899	401/231/632	0.840
TS + margin	0.3	0.900	396/232/628	0.835

Table 1: *Accuracy and filter types*

Next, we investigate how the accuracy changes as big posts are filtered out as foggy blogs. At first, for each blogger u of Doblog users, we group randomly selected his/her 30 blog posts. We here propose three new conditions to filter out foggy blogs followed by the filtering approach in Section 2. We call the filtering approach in eq. (2) as *small-filtering*, eq. (3) as *frac-filtering*, and eq. (4) as *margin-filtering*.

$$s^u_f \le c_s \quad \text{or} \quad s^u_m \le c_s \tag{2}$$

$$|\log\left(s^u_f / s^u_m\right)| \le c_n \tag{3}$$

$$|s^u_f - s^u_m| \le c_n \tag{4}$$

Results of accuracy for parameters c_n and c_s are shown in Table 3.3. The figures in 'Male' and 'Female' columns show the number of big blogs estimated as male or female respectively. The 'Param.' column means c_n or c_s depending on filter types, and the figures show the tuned parameters for accuracy of around 0.85 and 0.90.

In conclusion, our approach for gender estimation achieves 85% accuracy for 92% bloggers (692/752) and 90% accuracy for 84% bloggers (632/752).

4. Conclusions

We used SVM and lexical features to estimate bloggers' gender from their blog posts, and achieved 90% accuracy for 84% bloggers. We also extracted gender related words from trained linear SVM. As a future work, we will further investigate to point out the similarity or difference between gender classifiers for English corpora and for Japanese corpora.

Acknowledgments

Doblog corpus and questionnaire survey data are provided by NTT Data Corporation and Hottolink, Inc.

References

[1] J. Brank, M. Grobelnik, N. Milić-Frayling, and D. Mladenić. Feature selection using support vector machines. In *Proc. of the 3rd Int. Conf. on Data Mining Methods and Databases for Engineering, Finance, and Other Fields*, pages 84–89, September 2002.