Automatic Estimation of Bloggers’ Gender

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

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

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

Abstract

Blogs have become a major force in our daily lives across the globe. Bloggers are using blogs for expressing their comments, opinions or thinkings for daily events, accidents, products, services etc. By aggregating others’ blogs, we can see the trend emerging on blogosphere in near real time. The trend is valuable for business people in establishing marketing maneuver. However, such blog based approaches face a serious problem: trend for men and for women are definitely different in most domains. 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 some approaches employing Support Vector Machine (SVM) to estimate bloggers’ gender from blog posts as well as to extract gender related words. The data we use consists of blog posts on Doblog (Japanese blog-hosting service) and questionnaire results by Doblog users. Experimental evaluations show that our approaches achieve 85% accuracy for 92% bloggers and 90% accuracy for 83% bloggers as well as obtain gender related words.

Keywords

Gender Estimation, Text Classification, User Profiling

1. Introduction

Blogs have become a major force in our daily lives across the globe. Japanese Ministry of Internal Affairs and Communications released that the number of Japanese bloggers reached 8.7 million (it covers about 7% of Japanese population) in the end of March 2006. Bloggers are using blogs for expressing their comments, opinions or thinkings for daily events, accidents, products, services etc.

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 establishing marketing maneuver. For example, 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 some approaches employing Support Vector Machine (SVM) to estimate bloggers’ gender from blog posts as well as to extract gender related words. The data we use consists of blog posts on Doblog\(^1\) (Japanese blog-hosting service) and the result of questionnaire survey for Doblog users.

This paper is structured as follows. In Section 2, we give an overview of the related work on gender estimation from documents. Then, we propose our approaches of gender estimation and to extract gender related words from blog posts in Section 3. We show experimental results and discussions in Section 4, and conclude this paper in Section 5.

2. Related Work

Many works proved that text classification approaches employing Support Vector Machine (SVM)\(^3\) achieve high performance than other statistical approaches. For example, T. Joachims showed the superiority of SVM by comparing Naïve Bayes, Rocchio, C4.5, and k-NN\(^7\).

As for gender estimation, some works were trying to estimate author’s gender from documents. J. Holmes distinguished characteristics of male/female linguistic styles\(^6\). P. Excert targeted spoken language\(^4\), M. Palander analyzed electronic communications\(^10\), and S. Herring investigated correspondence\(^5\). J.A. Simkins asserted that there are no difference between male and female writing style in formal contexts\(^12\). M. Koppel et al. estimated author’s gender using the British National Corpus (BNC) text\(^8\). This corpus includes 920 documents, each of which includes 34,000 words on average, and author gender and genre are labeled for each document. 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. M. Corney et al. estimate author’s gender from e-mail content\(^2\).

In addition to function words and part-of-speech n-grams as in\(^8\), they used HTML tags, the number of empty lines, average length of sentences and so forth for features for SVM. They finally achieved about 70% classifi-
3. Gender Estimation

3.1 Overview of Proposed Approach

Our proposed approach consists of four phases: preparation phase, training SVM phase, filtering phase, and classification phase as follows.

- In preparation phase (see Section 3.2 in detail), feature vectors are created from blog posts.
- In training SVM phase (see Section 3.3 in detail), we train linear SVM and non-linear SVM from prepared dataset in Section 3.2.
- In filtering phase (see Section 3.4 in detail), foggy blog posts are filtered out by using the weights of features obtained from trained linear SVM.
- In estimation phase (see Section 3.5 in detail), bloggers’ gender is estimated by applying non-linear SVM.

We also show an approach of extracting gender related words in Section 3.6

3.2 Preparing Training/Test Dataset

The training/test dataset for SVM is prepared from blog posts provided by Doblog. At first, we apply morphological analysis to blog posts in advance to insert spaces between words because of no space between them in Japanese sentences. Then, 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.

3.3 Training Linear/Non-Linear SVM

For the training dataset prepared in Section 3.2, we train linear SVM and non-linear SVM respectively.

Equation of linear SVM is shown as

\[
prediction(x) = \text{sgn}[b + w^T x],
\]

where \(w\) is a weight vector, \(x\) is a feature vector, and \(b\) is a bias variance. Non-linear SVM is an extension of linear SVM with a kernel function. Equation of non-linear SVM is as follows.

\[
prediction(x) = \text{sgn}[b + \sum_{i=1}^{m} w_i K(x_i, x)]
\]

\[
K(x_i, x) = (1 + x_i^T x)^p
\]

where \(x_i\) shows a support vector and \(m\) is the number of them. \(K(x_i, x)\) is called kernel function. Among different types of kernel functions, we choose Polynomial kernel function with \(p = 3\) (three dimensions). We use SVMlight software\(^2\) for linear and non-linear SVM. The details of SVM are out of scope of the paper. See [3, 13] for details.

3.4 Filtering Out Foggy Posts

3.4.1 Male, Female, and Neutral

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’. Foggy posts can be classified into two types.

- Confusable post: The blog post that has both male features and female features at the same degree. Confusable post can be identified since it is located very close to the dividing hyperplane found by SVM.

- Short post: The blog post that lacks the proof for gender estimation. Short post can be identified because of few features.

\(^2\) http://svmlight.joachims.org/
By filtering out foggy posts, the precision of classification is expected to improve although the recall would be down.

### 3.4.2 M-Score and F-Score

To identify foggy posts, we propose ‘M-Score’ and ‘F-Score’ each of which shows the plausibility of the gender of the blogger as male or female respectively. M/F-Score is based on the weights of features for each feature. J. Brank et al. introduce one feature selection method where weights of features are obtained by using linear SVM [1]. In this paper, we employ their method for extracting weights of features.

The process of weight extraction is as follows.

1. Prepare input \( x \) for trained linear SVM.
   \[
   x = (x_1, x_2, \ldots, x_d) \quad (9)
   \]
   \[
   x_i = \begin{cases} 
   1 & \text{if } i = t \\
   0 & \text{if } i \neq t 
   \end{cases} \quad (10)
   \]

2. For an input \( x \), a score \( b + w_i \) is obtained as output (prediction(\( x \)) in eq.(6)). Note that we trained linear SVM with \( b = 0 \) in this study because of the simplicity.

3. The output \( b + w_i \) means the weight of feature \( t \), and the polarity – ‘positive’ or ‘negative’ – corresponds to the blogger’s gender. In this study, positive and negative features correspond to female and male respectively since we define positive example as female and negative example as male.

The weights for all features can be obtained by preparing all kinds of input \( x \). Here, we define M-Score, \( s_m^b \), and F-Score, \( s_f^b \), for a blog post \( b \) as follows.

\[
W_b = \text{a set of features in a blog post } b \quad (11)
\]
\[
w(t) = \text{the weight of feature } t \quad (12)
\]
\[
s_f^b = \sum_{t \in T_f} w(t) \quad (13)
\]
\[
s_m^b = \sum_{t \in T_m} w(t) \quad (14)
\]
\[
T_f = \{ t | t \in W_n, w(t) > 0 \} \quad (15)
\]
\[
T_m = \{ t | t \in W_n, w(t) < 0 \} \quad (16)
\]

#### 3.4.3 Filtering Out Foggy Posts

Confusable post can be identified by using M-Score and F-Score because its M-Score and F-Score are much the same. We define a parameter \( c_n \) for filtering confusable posts.

\[
| \log \left( \frac{s_f^b}{s_m^b} \right) | \geq c_n \quad (17)
\]

Short posts can be also filtered by M-Score and F-Score, because these posts have less features. The fewer the features in a post are, the smaller both M-Score and F-Score of the post are. We define the the parameter \( c_s \) for filtering short posts.

\[
s_f^b \geq c_s \text{ or } s_m^b \geq c_s \quad (18)
\]

By applying these two filters (eq.(17, eq.(18)) to new blogs as test dataset, we can filter out foggy blogs.

### 3.5 Process of Gender Estimation

Gender estimation process goes as follows. First, we train linear SVM and non-linear SVM with training dataset prepared in Section 3.2 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 as in Section 3.4. 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.

#### 3.6 Gender Related Words

Gender related words can be extracted from trained linear SVM as well as gender estimated blog posts.

The former approach is based on the weight and polarity of features (words) as described in Section 3.4.2. The weight of a word shows the degree of specificity of word usage, and the polarity of a word shows the gender using it. As positive examples are defined female and negative are male, words with positive weights represent female specific words, and words with negative weights indicate male specific words. By sorting positive/negative words with their weights, we can obtain the ranking of gender related words. We show an result by this approach in Section 4.2.

The latter approach is to employ tfidf score for gender estimated blog posts. Words with high tfidf in male estimated blog posts are considered to be male related words, and the same manner can be applied to female estimated blog posts. We employ this approach in Section 4.6.

### 4. Experimental Evaluations

#### 4.1 Training/Test Dataset

In this study, we use a collection of blog posts on Doblog and questionnaire results for Doblog users, both of which are provided by Doblog. The collection composed of 241,251 blog posts written in Japanese (157,817 males’ blog posts versus 83,434 females’ blog posts). The questionnaire results include Doblog users’ gender. As the blog posts and Doblog users are corresponding to each other, we can tell the gender for each blog post.

In preparing training/test dataset, 1,000 posts 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).
4.2 Feature Selection

We prepared three sets of training/test dataset depending on words’ part of speech.

Noun: A dataset consists of Noun only.

NAV: A dataset consists of Noun, Adjective and Verb, because these part of speeches are regarded as self-sufficient word in Japanese.

NAV+Tm: A dataset consists of “NAV” and the words in the end of sentence such as auxiliary verb because such words express the modality of sentences in Japanese.

Note that the words with low frequency ($\sum_p tf(t, D) \leq 3$ in Section 3.2) are removed from the features above. Then weights for features are obtained as in Section 3.4.2.

The accuracy of gender estimation for three sets of features is shown in Table 1. We can see from Table 1 that the performance is almost the same for any types of features. If we focus only on gender estimation, we should use Noun feature set because the smallest number of features will ease computational cost of linear/non-linear SVM. However, we hereafter use NAV feature set because extraction of gender related words is also one of our motivations.

The top 10 features of NAV feature set by positive/negative weights are shown in Table 2. From the Table 2, we can see that Japanese bloggers tend to use different expressions for the first person pronoun.

4.3 Accuracy for Varying Parameters

Next, we change filtering parameters, $c_n$ and $c_s$, for filtering foggy posts as described in Section 3.4.

Figure 2 shows the accuracy and the number of non-neutral posts with varying $c_n$ for filtering out confusable posts. If $c_n$ gets increased from 0 to 0.5, the accuracy are improved and the number of non-neutral posts are reduced. For $c_n > 0.5$, the accuracy are worsened although the number of non-neutral posts are also reduced.

Figure 3 shows the accuracy and the number of non-neutral posts with varying $c_s$ for filtering short posts. In contrast to the case of $c_n$ above, the accuracy is improved and the number of non-neutral posts are reduced consistently, if $c_s$ is increased at any value.

At last, we investigate accuracy and the number of non-neutral posts for varying parameters $c_n$ and $c_s$ simultaneously. Figure 4 plots the number of non-neutral posts and the accuracy (represented as F-value in the figure) for each pair of $c_n$ and $c_s$. This figure clearly represents the fact that the more neutral posts are filtered out, the more accuracy is improved.

4.4 Parameters Tuning

In this section, we explore two parameters $c_n$ and $c_s$ to make the accuracy highest. Under the condition of fixed F-value, there are many pairs of $c_n$ and $c_s$, depending on which the
number of non-neutral posts is different. These relations are shown at Figure 5 and Figure 6.

Figure 5 shows an inversely proportional relation between $c_n$ and $c_s$ under the condition of F-value being 0.8. As $c_n$ and $c_s$ have inversely proportional to each other, one parameter becomes smaller if other parameter becomes larger. Figure 6 shows a bell-shared relation between $c_n$ and the number of non-neutral posts under the condition of F-value being 0.8. If $c_n$ is larger than the peak of the curve, the number of non-neutral posts is reduced accordingly.

Small number of posts is filtered out for smaller $c_n$ as in Figure 6, however $c_s$ becomes larger accordingly as in Figure 5 and a large number of posts are filtered out for larger $c_s$. To keep more blog posts as many as we can, we have to set suitable parameters simultaneously. In the case of F-value being 0.8, $c_n = 0.24$ and $c_s = 5.55$ are obtained as suitable parameters.

4.5 Gender Estimation for Each User

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 the Doblog corpus including 241,251 blog posts, there are 752 users (480 male users and 272 female users), and a user posts about 320 blog posts on average. In this section, we use this corpus for gender estimation for a set of blog posts by the same user.

Figure 7 shows 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 prepare at least 30 blog posts when estimating the author’s gender.

Next, we investigate how accuracy changes as foggy users are filtering out. At first, for each of all users in Doblog corpus we prepare randomly selected 30 blog posts of his/her own as new test dataset. Then, we define M-Score, $s_m$, and F-Score, $s_f$, for a user $u$ as follows.

$$W_u = \text{a set of features in user } u$$ \hspace{1cm} (19)

$$w(t) = \text{the weight of feature } t$$ \hspace{1cm} (20)

$$s_f^u = \sum_{t \in T_f} w(t) \times tfidf(t, D)$$ \hspace{1cm} (21)

$$s_m^u = \sum_{t \in T_m} w(t) \times tfidf(t, D)$$ \hspace{1cm} (22)

$$T_f = \{ \forall t \in W_u, w(t) > 0 \}$$ \hspace{1cm} (23)

$$T_m = \{ \forall t \in W_u, w(t) < 0 \}$$ \hspace{1cm} (24)

Here we name M/F-Score by eq. (11)~(16) as ‘DS-score’, and M/F-Score by eq. (25)~(24) as ‘TS-score’ to avoid confusion. Note that the main difference between DS-score and TS-score comes from the embedded tfidf$(t, D)$ score.

We also propose three approaches to filter out foggy users as in the same manner as filtering approach by eq. (17) and (18) in Section 3.4.3. We call the filtering approach by eq. (25)
basic verbs as well as the gender related words related to ‘family’, ‘eat’, and some different words are used in male/female users. Table 4 shows we show some examples about specific topics to know how appears in female posts and ‘kin’ appears in male posts. This disproportionately without reasonable cause, i.e., ‘relative’ pear in female posts. However, ‘relative’ and ‘kin’ appear mainly appears in male posts, ‘boyfriend’ and ‘mother’ appeared in female blogs. It is also natural to understand that ‘girlfriend’ blog that many family related words are obtained in Japanese English blog [11]. We also find the same tendency in Japanese estimated blog posts.

4.6 Extraction of Gender Related Words

Once we obtain male/female estimated blog posts, gender related words can be extracted by $tfidf$ score. In this section, we show some examples about specific topics to know how different words are used in male/female users. Table 4 shows the gender related words related to ‘family’, ‘eat’, and some basic verbs as well as $tfidf$ scores extracted from gender estimated blog posts.

J. Schler et al. reported that female authors are apt to use family related words more frequently than male authors in English blog [11]. We also find the same tendency in Japanese blog that many family related words are obtained in Japanese female blogs. It is also natural to understand that ‘girlfriend’ mainly appears in male posts, ‘boyfriend’ and ‘mother’ appear in female posts. However, ‘relative’ and ‘kin’ appear disproportionately without reasonable cause, i.e., ‘relative’ appears in female posts and ‘kin’ appears in male posts. This might show our priori or posteriori nature specific to gender.

The usage of ‘eat’ has little difference between male and female, however some phrases composed of ‘eat’ have a big difference. For example, ‘eat all’ and ‘eat cake’ have high score in female blog posts than that of male. On the other, ‘be ate’ and ‘overeating’ have high score in male blog posts than that of female. From these tendencies, we might guess that female loves a cake and male might feel concern about obesity or metabolic syndrome.

As for other verbs, there is no significant difference between male and female. However, ‘run’ and ‘go out’ are dominant for female and ‘buy’ is for male. From these disproportionate usage of words, we can see the general preference derived from gender.

5. Conclusions

In this paper, we proposed an approach to estimate bloggers’ gender as well as to extract gender related words from their blog posts.

We all know from our daily experience that products, services, foods, health, etc should be designed, customized, and optimized for target gender because our behavior, preference, thinking, and recognition are somehow different between male and female. The proposed approaches, therefore, will have a great impact on a business domain because it will give us an opportunity to know gender related information. In fact, proposed approaches have been implemented on ‘Dentsu Buzz Research’[^3], a commercial online marketing service since November 30th 2006.

The questionnaire result we use in this study includes other information such as users’ age and residential area. In the future work, we will further investigate age estimation and residential area estimation to deeply understand user profiles from the archives of blog posts.

[^3]: [https://www.dbuzz.jp/](https://www.dbuzz.jp/)
Acknowledgments

Doblog corpus is provided by NTT Data Corporation and Hottolink, Inc. The questionnaire survey for Doblog users was conducted by Prof. Asako Miura and Prof. Satoshi Kityayama.

References


