Weblog Analysis and Classification

Puneet Chhabra
iCME, Stanford
pun@stanford.edu

Srinidhi Kondaji
EE, Stanford
skondaji@stanford.edu

Nagarajan
iCME, Stanford
nag.rajan@stanford.edu

December 12, 2008

Introduction

The text content of a large set of weblogs was analyzed along with its associated metadata (containing information related to the URL, author's gender, age, interests, location, etc) to classify blogs based on the gender and the age of the author using machine learning techniques. In particular, the following approaches were investigated:

1. Multinomial Naive Bayes Classification.
2. Support Vector Machines using LibSVM.
3. Feature Improvement through χ^2 analysis.

Data-set Preprocessing

A data-set of about 3 million blogs was obtained from Prof. Sepandar Kamvar(iCME, Stanford), and around 65000 english language blogs were filtered out from *Blogger* and *LiveJournal* which contained the complete author information including age and gender.
The text from each of these blogs was extracted using Python and an initial dictionary was created which contained around 140,000 words, with help from WordNet dictionary. During the extraction, a standard list of stop words, which occur frequently but have low classification value (a, an, the, of, etc), were eliminated and similar words were stemmed to a common root. From this initial dictionary, words occuring rarely were eliminated on the assumption that they would have little information about the author, and most of them were products of typing errors and occurred only once or twice in the whole list. This resulted in a set of around 35000 unique words in the vocabulary.
Using this dictionary of 35000 words, the blogs were converted into feature vectors for small storage and fast retrieval. These feature vectors were subsequently used for training the Naive Bayes and the SVM classification algorithms.

Implementation

The Python programming environment was used due to its strong text processing capabilities. The WordNet dictionary Python port was additionally used for recognizing words from the English language as well as for stemming purposes.

Stemming and Vocabulary

Instead of only using the Wordnet dictionary stemming function (*morphy*) we implemented a combination of Porter stemming algorithm along with the Wordnet dictionary stemmer. This gave a better stemming result than just using either of the methods.

After filtering low frequency words from this stemmed vocabulary, blogs were converted into variable length vectors. Figure 1 shows the relative frequency (log scale) of the various extracted

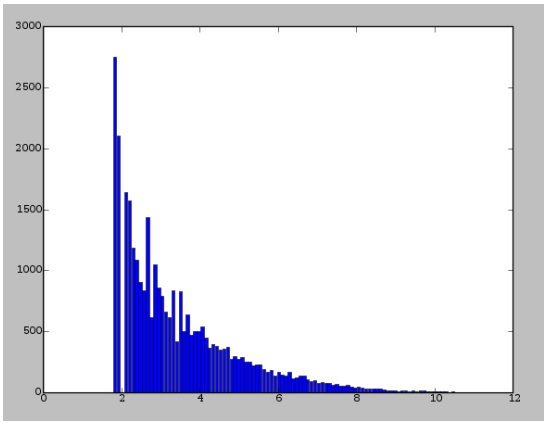

Figure 1: Relative frequency of features

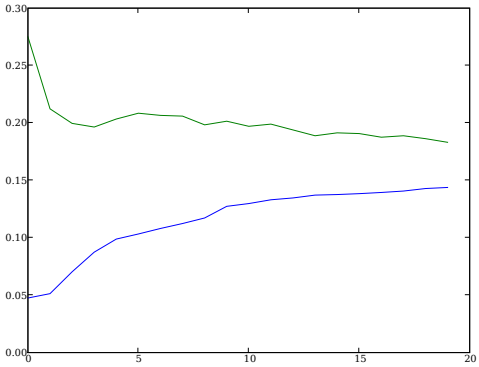

Figure 2: Learning Curve for Gender Classification (Naive Bayes)

stemmed words.

Results

Gender Classification using Naive Bayes

The first approach we tried was to implement a multinomial Naive Bayes Classifier using the set of 30,000 words in the vocabulary as the features. The sample complexity was found by plotting the learning curve which is shown in Figure 2.

The data was then divided into 66 sets of 1017 blogs each. To estimate the generalization error, k-fold cross validation was performed by leaving out one set of blogs at a time. The generalization error was found to be 24%. It is interesting to note the words which were most indicative of the author's gender (Table 1).

Male	Female
Mathematician	Charmer
Radeon	LivingRoom
CNet	BabySit
Engadget	HomeSchool
Jihadist	Popsicle
Information	Swarovski
Parakeet	Pediatrician
GeForce	Anthropology
Democracy	Petticoat
SourceForge	MySpace

Table 1: Top Distinguishing Words

As can be seen from the learning curve (Figure 2), the naive bayes classifier has a high bias. To reduce the bias, the complexity of the model was increased using a SVM classifier.

Gender Classification using SVM

LibSVM was used for SVM based gender classification, using a smaller set of features, to increase the speed of computation. A set of 5000 and 10000 most frequent words were selected as features. The learning curve for the the SVM classifiers using the above features are shown in Figures 3 and 4 respectively.

As can be seen, there was no improvement from the naive bayes approach. A χ^2 based feature selection method was implemented to enhance the classifier performance. Some of the top and the worst words as found by this method are shown in Table 2.

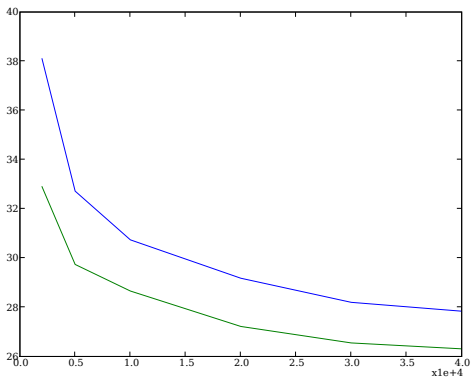

Figure 3: Gender Classification with SVM using 5k High-Frequency features

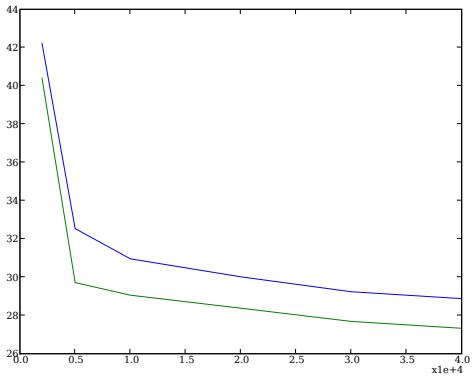

Figure 4: Gender Classification with SVM using 10k High-Frequency features

Best Words	Worst Words
Love	Quiznos
Obama	Dynasty
Google	Algebra
McCain	Shameless
Husband	Strangler
Internet	Hedonist
America	Hoodlum
Blogger	Astronomy
War	Nymphomaniac
Technology	Gravitate

Table 2: Words from Chi-Square Analysis

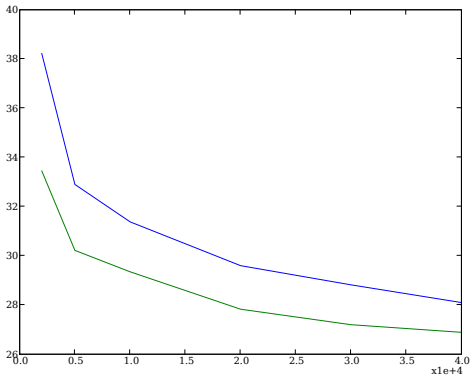

Figure 5: Gender Classification with SVM using 5k Chi-Square features

The learning curve for SVM classification using the best 5000 and 10000 words obtained from χ^2 based feature selection are shown in Figures 5 and 6.

Age Classification

The naive bayes approach was extended to classify blogs based on the author's age, first into 2 classes and then into 5 classes. In the first case, the two classes were chosen as the age groups (0-20, 50+) and in the second case the five classes were the age groups (10-25, 25-30, 30-35, 35-45, 45+). The five classes were chosen to have a uniform distribution of the number of blogs across the age groups.

The learning curves for the 2-class and 5-class age classification follow in figures 7 and 8 respectively. The 2-class age classification performed well with a hold-out test error of 15%.

Conclusion

Intuitively, age and gender classification based on text analysis is difficult, even for a human. So, it is not surprising to see that the performance was not as high as those observed in other text classification problems, like spam and document classification in news groups. However, the achieved ac-

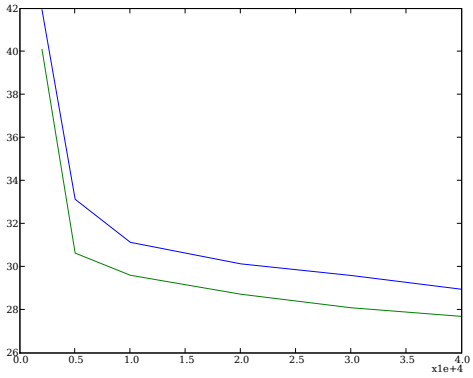

Figure 6: Gender Classification with SVM using 10k Chi-Square features

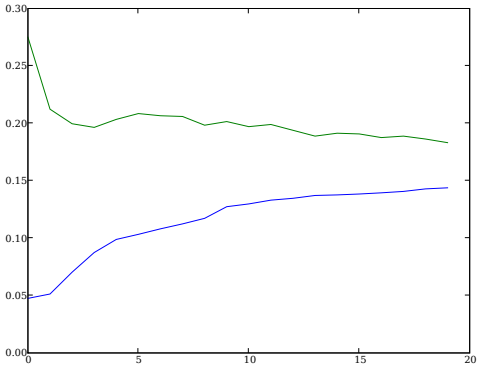

Figure 7: 2-class Age Classification using Naive Bayes

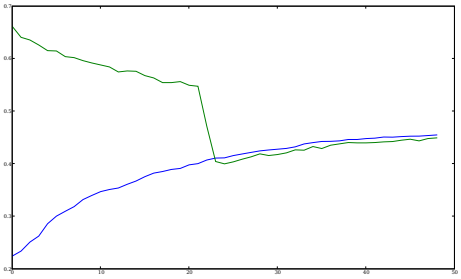

Figure 8: 5-class Age Classification using Naive Bayes

curacy is encouraging and our analysis show that there exist distinguishing features, across gender and age groups, which can be exploited using machine learning approaches.

References

[1] Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler, 2007, *"Mining the Blogosphere: Age, gender and the varieties of selfexpression"*, Text, volume 23, number 3, pp. 321.

[2] S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003, *"Gender, Genre, and Writing Style in Formal Written Texts"*, Text, volume 23, number 3, pp. 321.

[3] J.D. Burger and J.C. Henderson, 2006, *"An exploration of observable features related to blogger age,"*, First Monday, volume 12, number 9, "http://eprints.pascal-network.org/archive/00003406/01/".

[4] S. Herring and J. Paolillo, 2006, *"Gender and genre variation in weblogs,"*, Journal of Sociolinguistics, volume 10, number 4, pp. 439.