Mining Feelings

Luis Adarve-Martin, Jie Li, Eelan Chia

I. Introduction

Weblogs are personal, journal-style Internet entries that provide a glimpse into the blogger's life or his opinions regarding a matter of interest. Following its debut in 1996, the rapid growth in the number of blogs has also led to an increased interest by both the academic and business community. Some examples of ongoing research in this area include social network analysis and the mining of information and product opinions. Our research focuses on a combination of mood classification and the effects of age and gender on blogging; our goals include 1) the prediction of age and gender using feelings alone and 2) the clustering of feelings and location based respectively on co-occurrence in people and feelings.

II. Data Pre-Processing

The data for the project is taken from the website http://wefeelfine.org[1]. Presently, there are a total of 5,637 valid feelings present within this database. If a sentence within a weblog contains the phrase "I feel *X*" or "I am feeling *X*" and the feeling *X* corresponds to a valid entry, the sentence and the associated feeling is extracted from the database, together with other information such as the URL, author's age, gender, geographical location and weather conditions at the time of the weblog post.

Instead of using all the valid feelings in the database, we first reduced the list down to the most commonly occurring valid feelings. This was done by counting the number of times a valid feeling occurred in a Google query "feel *X*" or "feel like *X*", where *X* denotes the feeling. If a feeling occurred more than 5000 times in a Google query, the feeling was marked viable; this step resulted in a final list of 362 viable feelings. The next step was to use a script to obtain all the information (age, gender etc.) corresponding to the viable feelings. After we extracted the required information from the database, we clustered the information according to the URL. Thus, each unique URL is an entry in the extracted dataset and the values corresponding to a feeling are the total number of occurrences of the feeling in the weblog. For example, if a blogger expressed the feeling "happy" only on two occasions, the feeling "happy" will take the value of 2 for that entry. Given this implementation, a person with more than one weblog will thus appear as multiple separate entities.

The extracted data was found to be skewed and relatively noisy. Various research studies indicate that the majority of bloggers are teenage girls and that the mean age of bloggers is approximately 28 years. Our data substantiate this claim as most bloggers in our dataset fall between the age ranges of 10 – 30 and females bloggers account for 63% of all teenage bloggers (bloggers within the age of 10-20). (Refer Figure 1)

As we are interested in predicting the age and gender of a blogger, we next removed all entries that did not contain any of this information. We also removed entries in which the age listed was greater than 60. The likelihood of a person having a weblog decreases with increasing age and entries with an age value greater than 60 are likely to contain spurious age data.

---

[1] This website was chosen as we are working with Professor Sep Kamvar (one of the co-originator of the website) on this project and we inform him of the results obtained from this research.

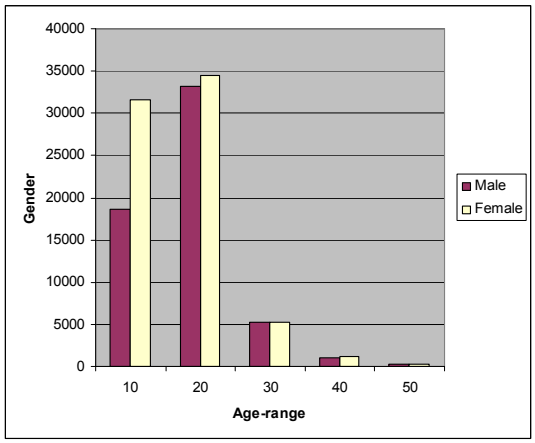

Figure 1: Plot of gender versus age range

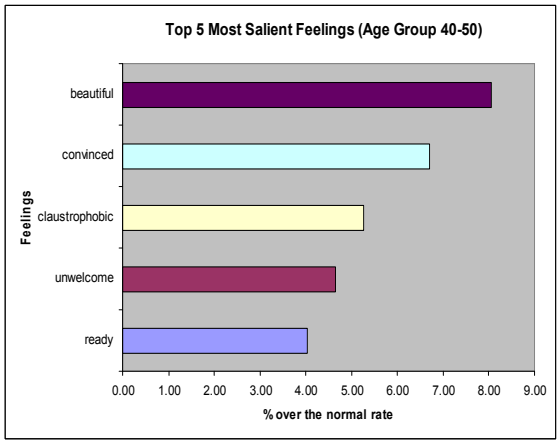

Figure 2: List of top 5 most salient feelings for the age-group 40-50.

After performing the pre-processing steps, each entry in the final dataset consists of a count of the number of occurrences of a feeling within a weblog, for all viable feelings, as well as information relating to either the age or gender and other information such as the weather condition and geographical location. A check of the data revealed that not many people expressed feelings within their weblog; as a result, the feature vector of the feelings was relatively sparse.

III. Feature Selection

Using this dataset, we next applied feature selection techniques to extract the feelings that allow us to distinguish between different age groups and gender.

We first examined the dataset to obtain the list of common feelings for each age group and gender. Examining the list of top ten common feelings by gender did not give us further insight as we found that the list was the same for both genders[2]. Next, we repeated the procedure for the list of salient[3] feelings. Amongst the 362 viable feelings, not many feelings appeared to be salient towards men or women. Similarly, there were not many salient feelings towards people in the different age-ranges of 10-20, 20-30 and 30-40. However, we came up with a number of interesting observations for bloggers in the age-range of 40-50 and 50-60. (Figure 2) For example, it appears that bloggers in these age-ranges tend to feel more claustrophobic as compared to the entire population. The proportion of people who expressed this feeling in their blog was 5 times over the normal rate (for age-range 40-50) and 10 times over the normal rate (for age-range 50-60).

Two feature selection techniques [4] were used to obtain the set of feelings that distinguished different populations - the χ^2 statistic and the information gain. The χ^2 statistic measures the lack of independence between a term t and a category c – a natural zero of the χ^2 statistic indicates independence of term and feeling. The given measure is

[2] The top five common feelings of both genders include *better, bad, good, right, sorry*.
[3] Feelings are salient to an age group / gender if they occur with higher frequency in the age group / gender than in the whole population.

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

where N: total number of documents; A: number of co-occurrences of t and c, B: number of occurrences of t without c; C: number of occurrences of c without t; D: number of times in which neither t nor c occurs.

Information gain is a common feature selection technique in machine learning applications. The information gain of term t is defined as:

$$G(t) = -\sum_{i=1}^{m} P(c_i)\log P(c_i) + P(t)\sum_{i=1}^{m} P(c_i \mid t)\log P(c_i \mid t) + P(\bar{t})\sum_{i=1}^{m} P(c_i \mid \bar{t})\log P(c_i \mid \bar{t})$$

Implementation of these two algorithms gave us two additional separate sets of feeling feature vectors for input into the classification algorithms. Each entry x_i in the feature vector corresponds to the number of times the i-th feeling occurs within a specific weblog. As mentioned previously, most of the entries in the feature vector are either 0's or 1's as not many feelings are associated with a specific weblog; when feelings are expressed within the blog, the same feelings typically do not occur multiple times.

IV. Comparison of Classifiers

We compared the performance of a number of classifiers in predicting age range and gender given a set of feelings. In addition, we applied the method of hold-out out cross-validation in our experiments, using 70% of the data for training and 30% for testing. The Naïve Bayes classifier with the multinomial event model was chosen as a baseline classifier. In addition, we implemented the classifier with Laplace smoothing so as to avoid the cases when the probabilities ϕ_k's are approximately zero, that is,

$$p(x_j = k \mid y = m) = \phi_{k|y=m} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \wedge y^{(i)} = m\} + 1}{\sum_{i=1}^{N} 1\{y^{(i)} = m\} n_i + |V|}.$$

Prediction of age-range using feelings

We are interested in predicting age-range (y) using feelings (x). Since there are five age-ranges, we expect a human to be able to predict the correct age-range by chance 20% of the time. For this section, we compared the performance of the Naïve Bayes classifier versus the Support Vector Machine (SVM) classifier. For the SVM classifier, we initially considered a regularization parameter C=1; subsequently, we used 1000 training examples to obtain a regularization parameter C=0.5.

Table 1 contains the classification results of the various classifiers; we note that the classification accuracy is similar for the two classifiers.

	Naïve Bayes classifier (all entries)	SVM, C=1	SVM, C=0.5
Classification accuracy	52.1%	51%	53%

Table 1: Test set classification accuracy

In addition, we also tested the performance of the Naïve Bayes classifier on different feature sets and sizes.

- Model 1: Prediction of age (y). Independent variables x: feelings.
- Model 2: Prediction of age (y). Independent variables x: feelings and gender

The classification results for this experiment are shown in Table 2.

	Model 1	Model 2
Naïve Bayes classifier (all entries with at least one feature expressed in weblog)	52.1%	52.6%
Naïve Bayes classifier (only entries with more than five feelings expressed in weblog)	52%	53.0%
Naïve Bayes classifier (only entries with more than ten feelings expressed in weblog)	45.4%	50.0%

Table 2: Test set classification accuracy for the Naives Bayes classifier (predicting age)

We note that the classification accuracy is lowest for the case when we considered only entries with more than 10 feelings expressed – this is due in part to an insufficient number of training examples with more than ten feelings in the feature vector. In addition, the results also indicate that knowing the gender does not significantly improve the classification accuracy.

Prediction of gender using feelings

Our next model of interest is to predict gender (y) based on feelings (x). Since there are only two classes, we expect a human to make the correct prediction by chance approximately 50% of the time. In this section, we considered the use of the Naïve Bayes classifier (with Laplace smoothing), decision tree and the Multi-Class Real Winnow (MCRW) algorithm [2].

	Naïve Bayes classifier (all entries)	Naïve Bayes classifier (only entries with more than 5 feelings)	Naïve Bayes classifier (only entries with more than 10 feelings)
Classification accuracy	56.5%	70%	90.5%

Table 3: Test set classification accuracy for the Naives Bayes classifier (predicting gender)

From Table 3, we observe that the Naïve Bayes classifier gives the best performance for the dataset in which individual feature vectors contain more than ten entries. However, as noted previously, it is difficult to find a big dataset that fit this criterion. Thus, we next compare the performance of the various classifiers using the dataset in which each feature vector contains more than five feelings. In addition, instead of using all 362 viable feelings, we consider only the features obtained using the χ^2 statistic. Results are shown in Table 4. We note that there is a slight improvement in classification accuracy for the Naïve Bayes classifier (79% as compared to 71% previously), thus the feature selection method was able to choose good (feeling) features that better distinguished the genders.

Entries with more than 5 feelings	Naïve Bayes classifier	Decision Tree	Multi-Class Real Winnow
Classification accuracy	79%	65.1%	52%

Table 4: Test set classification accuracy using only features obtained from χ^2 statistic

V. Secondary Analysis: Clustering and Co-occurrence

A secondary investigation involves the determination of similarity between different locations and the clustering of similar feelings. This was done using the similarity measure $Sim(x,y) = \frac{\sum_i x_i y_i}{\sqrt{(\sum_i x_i^2)(\sum_i y_i^2)}}$ and the co-occurrence matrix of feelings. In our experimentation, we noted that the countries that were similar to the United States, in terms of proportion of individual feelings expressed in weblogs, include Japan (similarity measure=0.98), Germany, France, Ireland (0.97) and Sweden (0.96). Another observation pertains to clustering of similar feelings. A person who expresses the feeling of "unwell" in one weblog is also likely to express the feeling of "better". The most commonly expressed feeling – "better" – also co-occurs frequently with feelings of "good", "bad", "right", "sick", "down" and "guilty".

VI. Conclusion

Given a set of feelings, we used feature selection to obtain those feelings which allow us to distinguish between different gender and age-ranges. The feature selection algorithm using the χ^2 statistic also returned a suitable set of features that allowed for improved classification accuracy. We also compared the performance of various classifiers in predicting gender and age-range based on feelings alone. In both instances, the Naïve Bayes classifier gives the best performance; in addition the classification accuracy improves with more feelings expressed within the feature vector. The second part of our analysis is dealt with in lesser detail in this paper as information regarding the characteristics of the population and the similarities or differences between various cities and countries are statistics pertaining directly to the data available on the website http://wefeelfine.org and as such, are useful mainly for publication on the website.

Acknowledgements

The authors gratefully acknowledge Professor Sep Kamvar's assistance and guidance during the brainstorming sessions and for pointing out relevant research directions. We also wish to acknowledge Professor Andrew Ng's advice with regards to feature selection.

References

[1] Burger, John D. et al., "*Barely Legal Writers: An Exploration of Features for Predicting Blogger Age*," Technical Paper, MITRE Corporation, December 2005.
[2] Schler, Jonathan et al. "*Effects of Age and Gender on Blogging*," AAAI Spring Symposium on Computation Approaches for Analyzing Weblogs, April 2006.
[3] Yan, Xiang et al., "*Gender Classification of Weblog Authors*," AAAI Spring Symposium on Computation Approaches for Analyzing Weblogs, April 2006.
[4] Yiming, Yang et al., "*A Comparative Study on Feature Selection in Text Categorization*," Proceedings of ICML-97, p. 412-420, 1997.