Chatous - Predicting the Quality of Users in Anonymous Chat

Team: Deepank Gupta, Aaron Li, and Simon Zhu

1. Introduction

Chatous is a text-based, 1-on-1 anonymous chat network that has seen 2.5 million unique visitors from over 180 different countries. Users can create a profile that contains a screen name, age, gender, location, and a short free-form "about me" field. After clicking the "new chat" button, users are matched up with one another in a text-based conversation. Interactions on Chatous include exchanging messages, sending/accepting a friend request, reporting an abusive user, ending a conversation.

One key challenge in anonymous chatting social network is to predict the user quality and likelihood to hold long conversation. Based on the user quality, Chatous can further generate better matching algorithms to pair people up in the anonymous conversation. In this report, we will use the datasets provided by Chatous including the user profile information and past conversation log to predict the quality of users and helps in predicting a better matching algorithm.

2. Raw Data Analysis

There are two datasets provided by Chatous:

1. 1.3 million user profiles including the following properties:
   |User ID| |Location| |Location Flag| |Age| |Gender| |Time created| |About| |Screen Name|
2. Logs of each conversation in Chatous platform containing the following information
   |Friendship Status| |Chat Created Date| |Chat Finished Date| Disconnector |Reported User ID| |Reason for Reporting| |First user Profile ID| |Second user Profile ID| |First user ID| |Second user ID| |Chat ID| |Length of Chat|

These two datasets include granular information about the quality of a conversation (e.g. length of a chat), demographic information of the chatter (e.g. age, gender and location), and information about the underlying network. They also contain the user profile ID, which makes it possible to run a panel data analysis to control users' heterogeneity. A key metric for Chatous is the intention to talk, which can be measured approximately by the number of lines in a conversation (Lines). By comparing the number of lines of two users we can infer which user has more intention to talk. The goal of this analysis is to build a model to make the prediction, based on the demographic information of the users.

We ran a k-means clustering algorithm on the profile data keeping age and gender as variables and summarize the statistics in Table 1.

Table 1: User Clustering Statistics keeping Age and Gender as Variables

Cluster Size	Medium Age	Gender
120776	24.5	Female
181799	11.0	Male
016416	66.0	Female
005464	42.5	None
001950	66.0	Female

Specifically males outnumber females in the network and the demographics also point to more teenage boys than girls in the networks.

Some other statistics about the user profiles based on their country are shown in Figure 1.

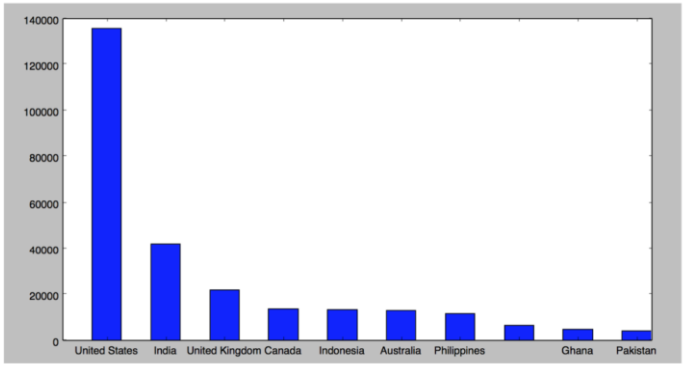

Figure 1: User Profile Statistics Based on Country

3. Model

3.1 User Classification

We got a hand-made manual set of 1034 users which were classified as either clean, dirty or bot. Using this data set and dividing it into training and test set; we tried to apply various classification algorithms in order to find out if we could learn to predict dirty/bot users and penalize them in the chat matching algorithm. We tried various different algorithms including

1. k-neighbors
2. Multinomial Naïve Bayers

3. Pipeline using feature reduction using Linear Support Vector followed by Random Forest classification.
4. Support Vector Machines using exponential kernel.

In the data set, the distribution is quite skewed; so instead of using the normal accuracy measure for calculating an algorithm's effectiveness, we used the Precision-Recall curves. We calculated the following metrics for every algorithm:

1. Precision recall curves
2. Number of false positives, false negatives, true positives, true negatives
3. Precision, Recall and F-score.

The AUC curves for various classification algorithms are shown in the figure 2 below.

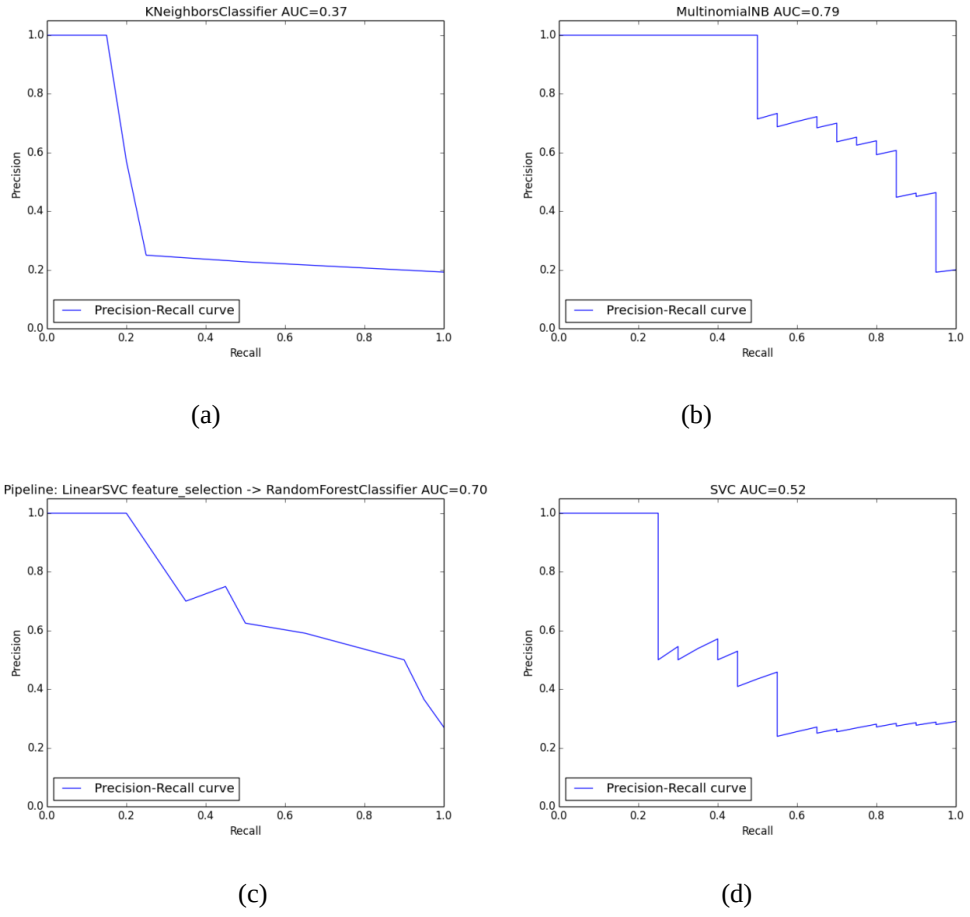

Figure 2: AUC curves for various classification algorithms: (a) k-neighbors, (b) multinomial Naïve Bayers precision recall, (c) pipeline precision recall, and (d) SVC precision recall.

From the figures above, we see that Multinomial Naïve Bayers precision recall has the highest AUC. Comparison of the detailed metrics for various algorithms is shown in table 2 below.

Table 2: Comparison of Classification Algorithms

	MultinomialNB	SVC	KNeighborsClasifier	Pipeline
AUC	0.79	0.52	0.37	0.70
Precision	0.62	0.55	0.57	0.75
Recall	0.75	0.30	0.20	0.45
F-Score	0.68	0.39	0.30	0.56
True Positives	15	6	4	9
False Positives	9	5	3	3
True Negatives	75	79	81	81
False Negatives	5	14	16	11

3.2 User Quality Regression

Besides the user classification, another key parameter in matching users is the quality of the user, which is defined as the average length of conversations that a user holds. If the user starts and maintains longer conversations, the user will get a better score. When we predict the quality of the users, we eliminated those who have not held a single conversation yet and we were thus left with 39727 users.

The parameter that we used for learning was the word vectors that the users had spoken till now. Based on the word vectors of their previous conversations, we tried to train a model with 70% of the user set and then use that to predict the quality on the rest of the 30% user base. We tried three different models:

1. linear regression
2. linear Support Vector machine
3. non-linear Support Vector Machine with an exponential kernel

The results were pretty much as expected that the non-linear support vector machine with exponential kernel performing really well. The results for the regression are shown in Table 3 below.

Table 3: Comparison of Regression Algorithms in Predicting User Quality (Length of Conversations)

	Linear Regression	Linear SVM	Nonlinear SVM
Mean Absolute Error	55.58	20.05	16.77
R2 Score	-55.57	-23.84	-0.06
Mean Squared Error	90692.83	39820.61	1692.11
Explained Variance Score	-55.57	-23.83	0.03

4. Analysis of Results

Using the various machine learning algorithms, we can build a better model which results in better matches by suggesting users with other higher quality users for chatting. We did two types of analysis: (1) classify users as legitimate clean users and (2) measure the length of conversation a user has as a proxy for user quality.

For the classification case, a hand-made dataset was used. As the dataset was small, algorithms like SVM which require a large dataset did not even perform as well as naive Bayes algorithm (as mentioned in Table 2). The ones that did the best were naive Bayes and feature selection followed by random forest technique. However, if we had larger dataset, we could do much better than the current estimations.

In the regression case, we obtained good quality estimates using non-linear exponential kernel for SVM. We also tried other models such as linear regression and linear kernel SVM but they did not perform as well as the exponential kernel for SVM. With the regression model, we could predict how likely a user is to keep up a conversation based on his/her earlier conversations. The dataset was again extremely sparse with nearly 50% of the sample users having done just 1 chat.

5. Conclusion

Using machine learning algorithms, we are able to get accurate interpretation on the demographics of the users with clustering. Also with the help of curated data sets, we built a convincing model to predict whether a user is clean or not. The regression model also helped predict user's inclination to talk based on their past conversational models even though the conversation logs are very few and feature space is very sparse.

Appendix: All the code for various algorithms can be retrieved in the following location: https://github.com/deepankgupta/chatous