Mining Facebook Data via Text, Images, and Likes*

Extended Abstract†

Wenfei Yin‡
Institute of Technology
Univ. of Washington: Tacoma
Tacoma, WA 98402
yinwf@uw.edu

Shreya Yembarwar§
Institute of Technology
Univ. of Washington: Tacoma
Tacoma, WA 98402
shreyay@uw.edu

Jonathan McFadden¶
Institute of Technology
Univ. of Washington: Tacoma
Tacoma, WA 98402
mcfaddja@uw.edu

ABSTRACT

KEYWORDS

ACM proceedings, LaTeX, text tagging

ACM Reference format:
Wenfei Yin, Shreya Yembarwar, and Jonathan McFadden. 1997. Mining Facebook Data via Text, Images, and Likes. In *Proceedings of TCSS 555: Machine Learning Term Project, Tacoma, WA USA, June 2018 (TCSS 555: Spring 2018),* 4 pages.
https://doi.org/10.475/123_4

1 INTRODUCTION

mnbmb

2 METHODOLOGY

Since there were three different types of data, we have divided describing our methodology into three parts. Each data type/source (*text, images, & likes*) will have its own part (*section*).

2.1 Text

For the Text part, we separate it into three parts. The first one is using the text to predict the accuracy of age and gender, then we use the 82- LIWC features to predict the big five personalities. For each part, we also use different methods to build the basic model to test accuracy.

We tried four different methods in text part, Linear regression, Random forest, Naïve Bayes and Logistical regression. Finally, we find the best result from the method which uses logistical regression and Naïve regression.

The first step of the text part is data preprocessing. In the files, we find that they are separated, thus we merge the profile file and text file according to the common primary key in each take. We add each text part into profile.csv. At the meanwhile, we also transfer the age to age groups, and we use 1 and 0 to replace the female and male in order to more effective.

After data preprocessing, we use the Naïve Bayes to predict the age and gender. As is known to us who focus on machine learning. Naïve Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes theorem with strong independence assumptions between the features. However, we find the model only working well on gender part. The accuracy of age in Naïve Bayes model is 0.49 which is lower than baseline (0.59). Therefore, we have to change our method. After researching, we guess the Logistical Regression will work well on age part. At the beginning, we split the training data into two part, the one is training data which are 8000 rows, the other is test data which is 1500 rows. To our surprise, it does not work well. After trying different methods, like linear regression, random forest. We find the model cannot work well when we use text to predict the age, thus we try to use 82-LIWC features to build model.

The LIWC, The linguistic Inquiry and Word Count tool, is known text analysis software which is widely used in psychology studies, In the file, for each user, it has 82 features. We have to merge the LIWC and profile file. When we create the model, we delete the big five personalities which need to be predicted in the table. After testing in local machine, we find it work well in age predict. The accuracy of age group reach to 0.62.

For Five Big Personality, we have to used 82-LIWC features. Because of big data, we decide to use Linear regression to set the model. In statistics, Linear regression is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables. Fortunately, Linear regression work well in the model, the results almost similar with baseline. Although we try to decrease the RMSE of five big personality, the methods we tried are almost same.

2.2 Images

jhljhv

2.3 Likes

When considering the Likes data, we first create a **User/LikeID Matrix** from the training data we are given. This process is described below in the Dataset and Metrics section. Based on the training data we were given, this resulted in a matrix with 536204 columns. Due to the large number of columns in this matrix, dimensionality reduction using Singular Value Decomposition or Principal Components Analysis was considered. Unfortunately, when the singular values of this matrix were computed (*up to the 9499th*

*Produces the permission block, and copyright information
†The full version of the author's guide is available as `acmart.pdf` document
‡
§
¶

TCSS 555: Spring 2018, June 2018, Tacoma, WA USA

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

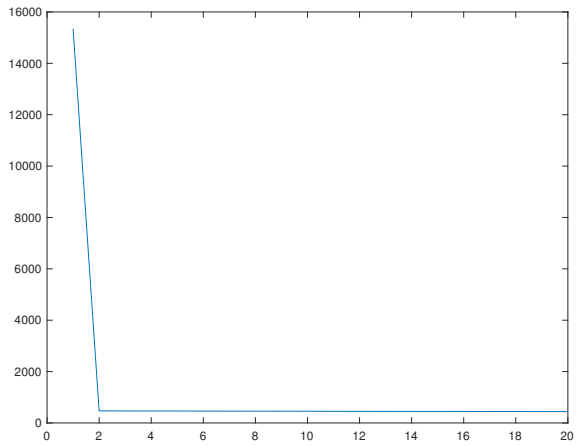

Figure 1: The type of 'cliff' drop-off in singular value magnitudes we wanted to see.

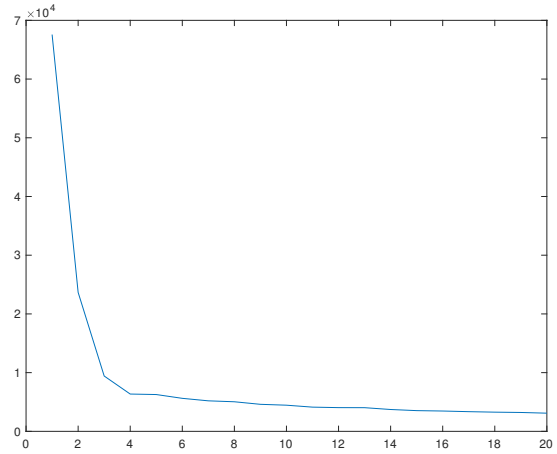

Figure 2: The singular value magnitudes we got instead of what we wanted to see.

singular value), there was no characteristic "cliff" drop-off in their magnitudes which would indicate a good point for truncation.

To illustrate this, the type of "cliff" in the singular value magnitudes we were seeking can be seen in the plot of singular value magnitudes, from a different dataset given in **Figure 1**.

Unfortunately, the result was got did not have the sudden drop over just a few singular values. What we got instead can be seen in **Figure 2**.

Thus, we proceeded to use the entirety of the column space of the **User/LikeID Matrix**. We constructed models using the machine learning algorithms

- Random Forest
- AdaBoost
- Bagging (*with in-bag scoring*)
- Bagging (*with out-of-bag scoring*)
- Naive Bayes
- K-Nearest Neighbor
- Support Vector Machine
- Linear Support Vector Machine
- Gradient Boosting
- Neural Networks (*implemented using Keras*)

Each of these algorithms was run using a variety of performance parameters, including the number of estimators, the learning rate, the tolerance, the number of neighbors, and more. Ultimately, for each of the user attributes, we chose the algorithms and settings

- **Age Group: Neural Network** with **750 base nodes** (*see below for full network design*
- **Gender: Neural Network** with **650 base nodes** (*see below for full network design*
- **OPE: SVM** with **tolerance of** 1×10^{-6}
- **NEU: AdaBoost** with **100 estimators and a learning rate of 1.0**
- **EXT: Random Forest** with **1500 estimators**
- **AGR: Bagging *using in-bag scoring*** with **250 estimators**
- **CON: Bagging *using in-bag scoring*** with **100 estimators**

We also strongly considered the alternative algorithms and settings

- **NEU: AdaBoost** with **250 estimators and a learning rate of 0.01**
- **EXT: K-Nearest Neighbors** with **500 neighbors**
- **AGR: K-Nearest Neighbors** with **800 neighbors**
- **CON: AdaBoost** with **1000 estimators and a learning rate of 1.0**

2.3.1 Neural Net Design. We designed three types of neural networks for this part of the project. The first type of neural net was a single-class binary classifier for predicting gender; while second type as a multi-class binary classifier for predicting age group. Finally, the last type of neural net we designed was a multi-layer linear regressor. All three networks had the same number of hidden and dropout layers, however they used different activation functions and kernel initializers. Additionally, all of these neural networks were implemented in **Keras** using its **Sequential** model. We summarize the three neural nets in **Table 1**.

3 DATASET AND METRICS

dadfsfa

Table 1: Likes SUMMARY of Neural Nets

	Gender	Age Group	Regressor
Input Layer			
Neurons	536204	536204	536204
1st Hidden Layer			
Dense	975	1125	750
Kernel	*uniform*	*default*	*default*
Activation	*relu*	*relu*	*relu*
1st Dropout Layer	0.25	0.25	0.25
2nd Hidden Layer			
Dense	1300	1500	1000
Kernel	*uniform*	*default*	*default*
Activation	*relu*	*softmax*	*relu*
2nd Dropout Layer	0.375	0.375	0.375
3rd Hidden Layer			
Dense	975	1125	750
Kernel	*uniform*	*default*	*default*
Activation	*sigmoid*	*sigmoid*	*sigmoid*
3rd Dropout Layer	0.25	0.25	0.25
4th Hidden Layer			
Dense	650	750	500
Kernel	*uniform*	*default*	*default*
Activation	*relu*	*relu*	*sigmoid*
Output Layer			
Neurons	1	4	1
Kernel	*uniform*	*default*	*default*
Activation	*sigmoid*	*softmax*	*sigmoid*

3.1 Text

There are five filed in the training dataset. LIWC, profile, relation and test. For text and LIWC part, we use the text folder which contain 9500 text files, and we merge text files and a csv file named "profile.csv" because of common primary key "userid". Meanwhile, we merge the profile and the LIWC file as well. In the profile, we use 0 denote male, and 1 is female. In the age, we have classified the age likes "xx-24", "25-34", "35-69", and "50-xx". In order to more effective, the big five personality, Openness, Conscientiousness, Extroversion, Agreeableness, Emotional Stability, being referred to as "ope", "con", "ext", "agr" and "neu".

3.2 Likes

3.3 Images

The image database contained 9500 images, which was split into 8000 for training and 1500 for validation. For the purposes of gender prediction, the 8000 images were regrouped according to the labels, i.e. 3386 males and 4614 females. The regrouping was done using the profile.csv file. One critical issue to note is that it is an unbalanced distribution between the two labels. Some of the issues noticed in the database were as follows:

- Images did not contain faces
- Images contained multiple faces
- Images did not contain face belonging to the label holder

Table 2: Text RESULTS for Age & Gender Accuracy

	Age Group	Gender
Baseline	0.59	0.59
Linear Regression	0.52	0.60
Random Forest	0.51	0.60
Naive Bayes	0.49	0.71
Logistical Regression	0.61^a	0.59
	0.59^b	

[a]Using LIWC data
[b]Using text data

Table 3: Text RESULTS for Personality Score RMSE

	OPE	NEU	EXT	AGR	CON
Baseline	0.65	0.80	0.79	0.66	0.73
Linear Regression	0.65	0.79	0.79	0.65	0.72

4 RESULTS

Our overall results can be summarized by the table .

Detailed results for the three different classes of data are given in their own sections below.

4.1 Text

For age prediction model, using Logistical regression based on 82-LIWC features is best choice in the method we tried. We also use other methods, line Linear regression, random forest, Naïve Bayes. We can see the results in the chart given in **Table 2**.

From the chart in **Table 2**, we can see that the best result in age prediction is the model when it uses logistical regression based on LIWC, and the best accuracy in gender predict model is Naïve Bayes. Therefore, our choice are Naïve Bayes based on text in gender and Logistical Regression based on LIWC in age prediction.

For Big five personality, we only use linear regression on the text and LIWC part, because we tried other methods, and the results are same. The results are in the chart given in **Table 3**.

From the chart in **Table reftab:txt-person**, we can clearly see that the results under the predict model are similar with Baseline when we use Linear regression.

4.2 Images

jhlgjhg

4.3 Likes

ljlkj

5 CONCLUSION

ljhgljhg

5.1 Text

For text part, we have tried many methods when we set up the model, but we do not have do more specific work on data preprocessing, such as feature selection, word choice. When we predict the age and gender, we can find more significant words, like she, he which will affect our results. Although we get a good result in the age prediction when we use Logistical regression based on 82 LIWC features, it is not perfect. In our last assignment, we find the neural network may be good at decrease RMSE in big five personality, thus we can try it in the future work.

5.2 Images

jbklj

5.3 Likes

bljk