Sajad, Raj, Alan
Team1 Group Report
March 10, 2022

User Profiling

Based on Text, Image, and Likes

1.0 Introduction

There is an abundance of user-generated content that is only increasing with each day. There is also a growing demand for personalized services, most common example being targeted advertisements. Studies in consumer psychology have revealed that many details such as age group and gender as well as personality traits can be determined from a picture of the target's face, their status updates, and the pages they like on social media. The goal of this study is to design a machine learning system that is capable of predicting the age, gender, and personality traits of a person given their social media profile pictures, status updates, and liked pages. In the project, we use a training set with 9500 labeled instances of Facebook users. We use accuracy score to measure the prediction performance of age and gender, and RMSE for personality. The results achieved are better than the baseline and have significant improvement compared to the very beginning.

2.0 Methodology

2.1 Image as a Source

The machine learning method implemented was transfer learning. It is the act of using an instance of a model that has been previously trained (preferably on large datasets for very long periods of time) for a task, for a different but similar task. The key here is that we intend to utilize the weights as these have been calculated during the model's initial training phase and take time to do so. We have used the famous VGG16 model, replacing the top layer as well as fully connected layers with one dense layer, one dropout layer with a dropout factor of 0.5, and finally

one sigmoid activation output layer which predicts among 2 classes as compared to the original 1000-class sigmoid activation output layer on the VGG16 model.

2.2 Likes as a Source

Several methods have been tested for likes-data, but the final version is based on the simple machine learning method: MNB; the feature selection method has been carefully selected. Our final version of prediction of age and gender with likes data is using MNB to predict age and data but we simply use a metric that pairs of [like_id, age] or [like_id, gender]. For each input user's likes, we simply predict age and gender for each single like id and use the result set to vote to a result. For dealing with the original data, a matrix which columns are selected likes ids, gender and age is generated.

For reducing feature space, we filter the like id that appear less than 50 times. We implemented several methods to fit that. They are MNB, SVM, logistic regression, CART and Adaboost. The result is shown in results section.

2.3 Text as a Source

There is no machine learning method that takes raw text data without re-presenting it in a numeric form. There are several methods to re-present text data. In this study we used two, namely Count Vectorizer and TF-IDF vectorizer. Those two vectorizer basically count words within a text based on different methodologies, and they provide the option to count by n-gram. This study utilizes this option to predict gender based on text.

Once the text data re-presented, we passed its representation into several machine learning models, namely SVM, MultinomialNB, Decision Tree, KNN, and Logistic Regression. As it is illustrated in the results section, the accuracy score in predicting gender significantly changed when we changed the data representation. This encourages us to do some educated modification into the dataset in order to change its representation to the machine learning models such as by eliminating non words from being counted by the vectorizer. However, this modification didn't impact the accuracy score significantly.

This modification to the representation strategy persists. We have tried it in the other form provided for text which is LIWC analyses (read section 2.4.3 Text). This form was mainly used to predict the big five personality traits. At the beginning, it was manual modification by eliminating columns that were obvious to be removed such as the userid column or 'Seg' column which contains only '1s' as its value. We have also removed a whole type of columns such as punctuations which are 5 columns. This manual approach was not successful until we used Recursive Feature Elimination method to recommend which columns to keep and which ones to remove from the representation. Using this method did indeed impact the accuracy score significantly but this impact is not guaranteed every time. The impact was measured by comparing between accuracy scores obtained using this method and accuracy scores obtained without using this method. The comparison shows that RMES score significantly impacted 3 out the 5 personalty traits we are predicting. This confirms that changing data representation do impact but this impact is not guaranteed every time.

2.4 Dataset and Metrics

2.4.1 Image :

There are 9500 images in .jpg format. The images are in non-uniform dimensions and are in RGB format. Each file has a title corresponding to the User ID of the person in the image. Categorical cross entropy loss was used as the error function during model training. The evaluation metric used is accuracy.

2.4.2 Likes:

More than one million records of users' likes relation data are taken into account as well as the 9500 users' profile data. For each likes record, there is only the mapping of user id and its like id.

2.4.3 Text:

The Text dataset provided in two forms: 9500 .txt files form, where each file named with a unique user id, and LIWC analyses form of these .txt files, where each row represent an analysis of one user. This dataset was used to predict a user's gender, which is a binary classification task,

and it was also used to predict a user's personality traits, which is a regression task. Therefore, the metrics used are suitable for these two tasks which are Accuracy score and Mean Squared Error respectively.

3.0 Results

3.1 Image as a Source

The model achieved a validation accuracy of 74% after 30 epochs and an accuracy of 72% on the actual test data during model evaluation. The complexity of the model was well-equipped to handle the various kinds of images since it had previously been trained to predict 1000 classes, so predicting 2 classes wasn't very difficult.

3.2 Likes as a Source

Table 3.2.1 Result of gender and age with likes data(separated)

	age_group	gender	remain likes	remain user
MNB	0.78	0.60	1871	7841

Table 3.2.2 Result of gender and age with likes data based on different methods

	age	gender
MNB	0.60	0.78
Logistic Regression	0.56	0.757
CART	0.58	0.67
ADABOOST	0.549	0.698
SVC(Linear)	0.55	0.73

Table 3.2.1 shows the final result that use data as the likes id, age and gender matrix; the remain likes are selected ones based on the selection with the frequency of more than 50. Table 3.2.2 displays the result based on different methods.

3.3 Text as a Source

Table 3.3.3 Results of predicting gender v1 Deployment-1

model	gender	representer
KNN	0.712	Count Vectorizer
Decision Tree	0.712	Count Vectorizer
MultinomialNB	0.712	Count Vectorizer

Numbers in the above table represents accuracy score metrics

Table 3.3.4 Results of predicting gender v2 Deployment-2

model	gender	representer	n-gram
LogisticRegression	0.7515	TF-IDF Vectorizer	1
	0.65	TF-IDF Vectorizer	2
	0.58	TF-IDF Vectorizer	3

Numbers in the above table represents accuracy score metrics

Table 3.3.5 Results of predicting personality traits Deployment-3

model	ext	neu	agr	con	ope
LinearRegression	0.65	0.60	0.42	0.48	0.38

Numbers in the above table represents Mean Squared Error metrics (RMSE)

Table 3.3.3 shows the first version of trying to predict gender. The observation is that changing the model does not change the accuracy score. However, there was a difference in processing speed. Decision Tree was slower. Tn Table 3.3.4, which is the second version of trying to predict gender, the representer was changed from count vectorizer into tf-idf vectorizer. This change immediately reflected into the accuracy score, and in our case, this change was positive as it went from .71 into .75.

Table 3.3.5 shows the first version of trying to predict big five personalty traits which are a continues value. Therefore, we used Linear Regression model to predict them. The green boxes indicates that those models were eventually deployed into our vm and evaluated in a different dataset that has similar structure.

4.0 Conclusion

The results achieved passed the baseline of 58% accuracy on the test data. Upon observation of accuracy and loss curves, it was understood that the numbers had stagnated at around the 30th epoch and by the 50th epoch, overtraining began to creep up. Ideally, training could be stopped early at the 30th epoch. Also, the images were resized, so further study into whether the images lost information during this process would be in order. The result of this would be an ideal format for the input images. We make use of several different machine learning methods to deal with the data of MyPersonality. From a text prospective, texts must be re presented, and the game seems to be in data representation more than it is in model selection. The model selection seems to be guided by other factors such as the types of output we are predicting or scalability requirements and speed, but it seems not guided by the nature of the dataset.