Personality prediction using Online Social Networks digital footprint

by Shahin Taghikhani

+1 (252) 565 - 6207
taghikhani.sh@gmail.com
Shahin-taghikani
STBlackHawk

Abstract

Understanding human behavior and personality is an essential phenomenon in the 21st century. Having the ability to predict a user or customer's personality has many applications in different domains and industries such as recommender systems, health care, e-learning, etc. This project implements a personality prediction system based on the Big Five personality model using text and pictures from social media Step by step implementation process of the project as below.

- Data Preparation
  - Platforms
  - Data sets
  - Labeling
- Modeling & Architecture.
- Data pre processing, train, development & test set.
- Evaluation metric, deliverables & error analysis.

Data preparation

Platforms:

- Twitter
- Flickr

Data sets:

- PsychoFlickr [1]

	Flickr		
Modality	**Total #images**	**Average #images per user**	**Median #images per user**
Posts	72,997	247	170
Likes	60,001	203	200
Profile Images	295	1	1

- Cross-linked Flickr-Twitter data set [1]

	Flickr		
Modality	**Total # images**	**Average # images per user**	**Median #images per user**
Posts	60,381	175	56
Likes	28,658	83	45
Profile Images	344	1	1
	Twitter		
Modality	**Total # images**	**Average # images per user**	**Median #images per user**
Posts	73,576	213	199
Likes	29,030	84	82
Profile Images	344	1	1

- TwitterText [2]

TwitterText			
Users	Tweets per user	Gender	Total tweets
66,502	3,200	31,307 Male & 35,195 Female	104,500,740

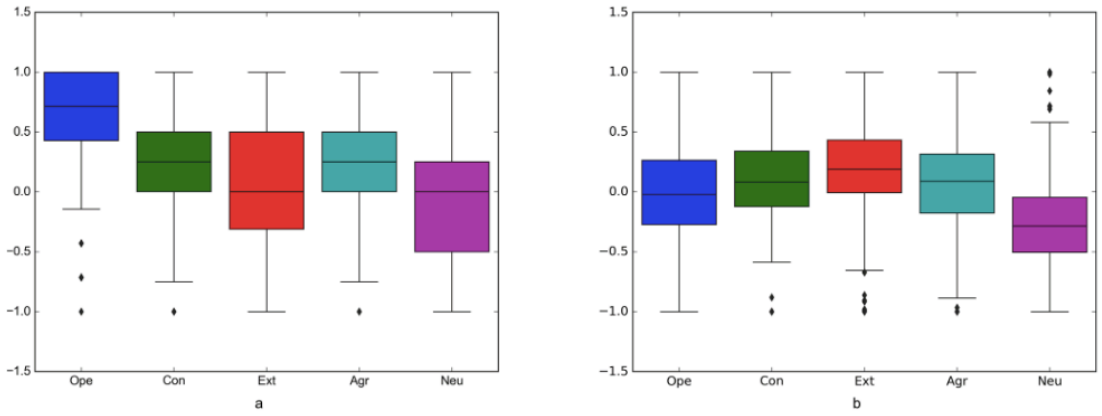
Labeling

- *PsychoFlickr:*
  This data set contains both self-assessed and perceived personality for each of 300 pro users of Flickr.
- *Cross-linked Flickr-Twitter data set:*
  This data set used an automatic text regression method to label the personality trait for each user. The model used in this method has been trained on samples over 70,000 Twitter users.

  Below shows the distribution of labeled personality traits for Big five model over each data set. Personality traits: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to experience

Distribution of different personality traits at the two data sets.

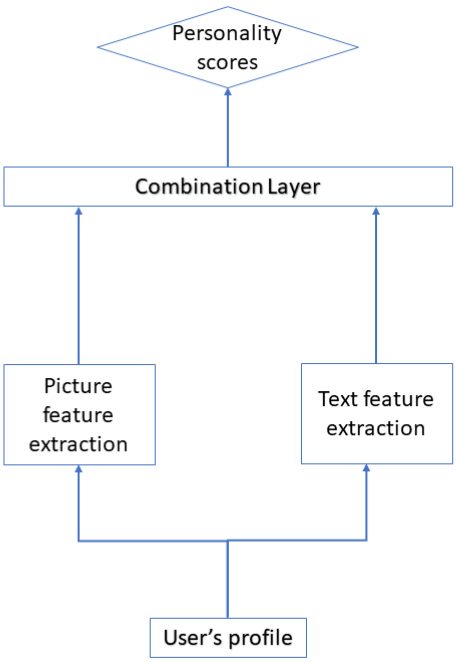
(a) Psycho-Flickr and (b) Cross-Linked Flickr and Twitter.



- *TwitterText:* For labeling, we use a method developed by [3] which they trained their model on a large sample of around 70,000 Facebook users who have taken Big Five personality tests and shared their posts using a model using 1-3 grams and topics as features.

Modeling & Architecture

The architecture consists of two parallel neural networks for feature extraction: One for text and the other for pictures. After extractions, features will be combined in a combination layer and then regression(s) such as Support Vector Regression, Gradient Boosting Regression, Random forest Regression & etc. will be used for final prediction score.
Below is the schematic of the whole model:



Text Text feature extraction is the implementation of [4]. Based on their work a model for deep learning based text posts will be implemented.
Implementation steps

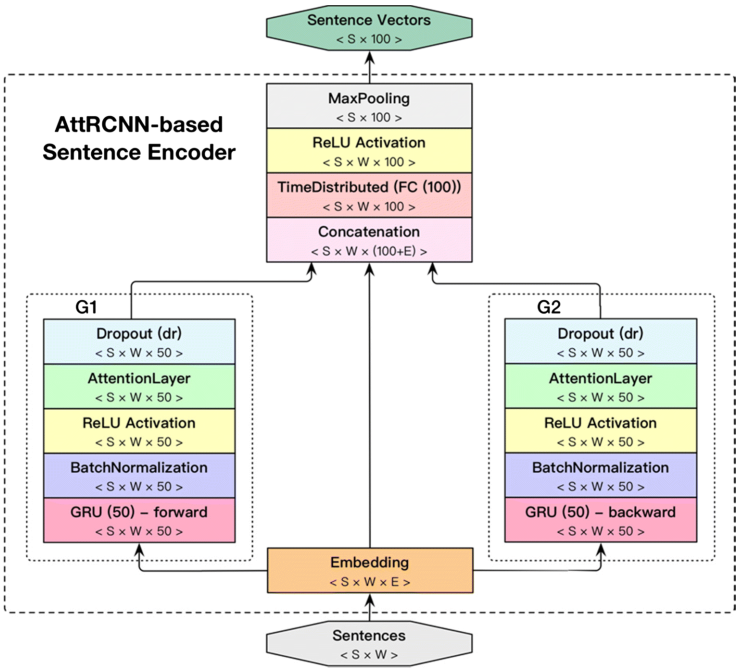
- Pre-train word embedding. CBOW model [5] & word2vec toolkit can be used for this step
- Construct the embedding matrix M
- Sentence vectorization using RCNN. As an example, AttRCNN-CNNs sentence encoder designed by [4] can be used.
- Document vectorazation using CNN.
- Feature extraction for combination layer input.

An example, a two-level hierarchical deep neural network model, named AttRCNN-CNNs [4] has been shown below. Furthermore a detail description of this network can be found in [4].

Designs • *Hierarchical neural network for text posts modeling* :

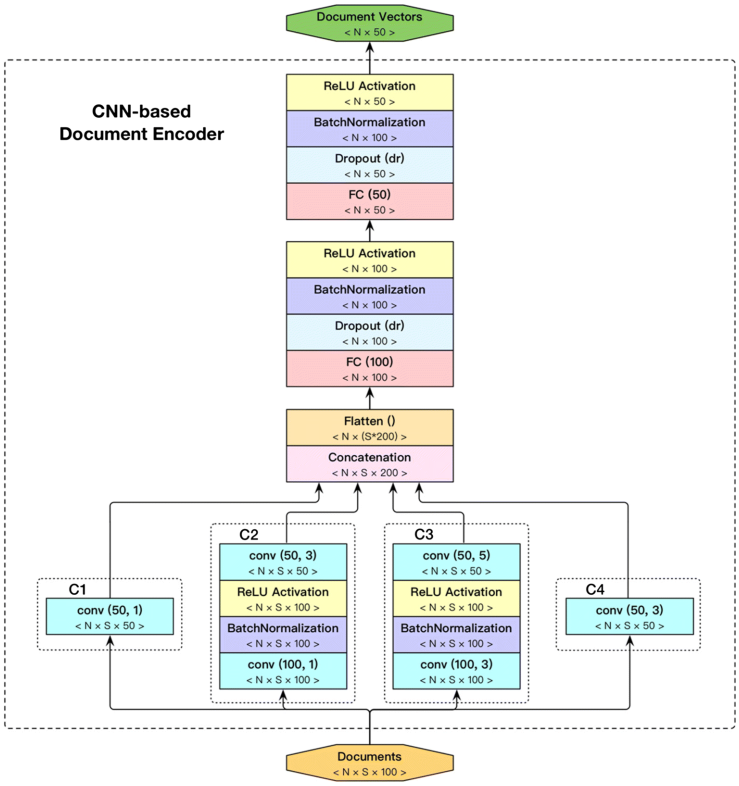
The fully-connected layer is donated as “FC (number of neurons)”. The shape of each object is shown within angle brackets, and so is the output shape of each layer

• *AttRCNN-based Sentence Encoder.*



The GRU layer and dropout layer are donated as “GRU (number of neurons)-scan direction” and “Dropout (dropout rate)”, respectively. The output shape of each layer is shown within angle brackets

• *CNN-based Document Encoder*



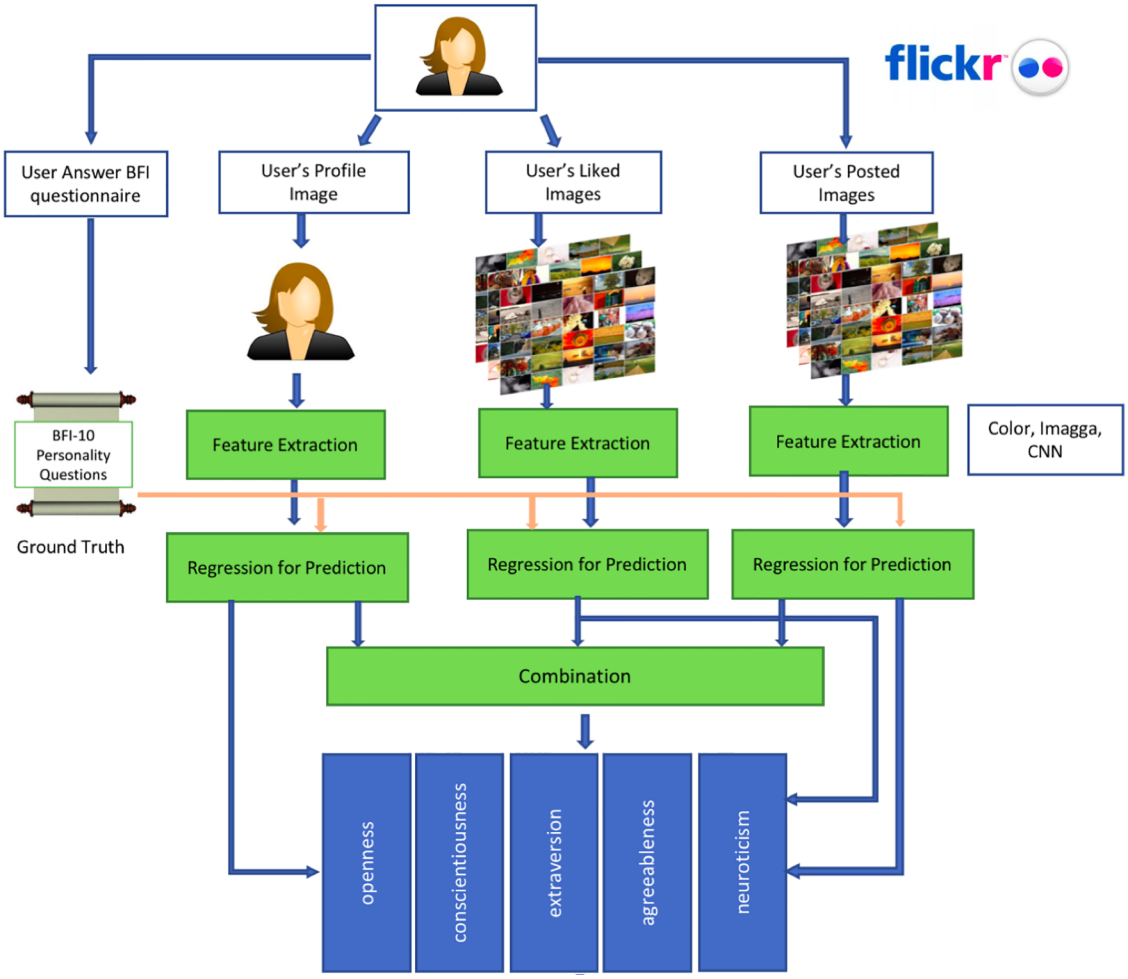
The convolutional layer, fully-connected layer and dropout layer are donated as “conv (number of filters, kernel size)”, “FC (number of neurons)” and “Dropout (dropout rate)”, respectively. The output shape of each layer is shown within angle brackets

Picture Feature extraction will be following the work of [1]. Based on various studies in this domain two categories of features can be extracted colors and content. For each user profile picture, features will be extracted and for liked and posted images per user mean feature pooling will be performed.
For content feature, VGGnet and Imagga tagging has been proposed by [1]

- *Description of features in image feature extraction [1]*

Feature Type	Dimension	Feature Name	Detailed Description
Color	1	Grayscale (binary)	if an image is grayscale or not. If the image is grayscale, then the rest of the features are not computed
	10	HSV statistics	Average and standard Deviation of hsv space, number of distinct hues, natural log of h_count
	12	Hue statistics	12 hue histogram (normalized, all 12 values sum up to 1)
	1	Pleasure (p)	Pleasure = 0.69∗Brightness+0.22∗Saturation
	1	Arousal (a)	Arousal = 0.31∗Brightness+0.60∗Saturation
	1	Dominance (d)	Dominance = 0.76∗Brightness+0.32∗Saturation
	6	6 Hue histogram	yellow, green, cyan, blue, magenta, red
Content features	1365	CNN object and scene probabilities	VGG_Net prediction on 1000 objects and 365 scene categories
	4096	CNN generic features	4096 dim penultimate layer features of VGG_Net
	1299	Imagga tags	list of Imagga tags for a set of images

- [1] Design for feature extraction can be used as an example



Data pre-Processing

Pre-processing step consists of pre-processing of texts including text tokenization and text unification, dividing the data into training, development and test set.

- *Text tokenization:* Adding necessary space between text elements (words, punctuation's, emoticons, URLs, numbers, etc.) and delete unnecessary spaces within a single text element, e.g., emoticons.
- *Text unification:* Reducing the length of tandem duplicated elements(busyyyy & busyyyyyyy) to make sure the length of such elements in a certain token is no more than 3.
- *Train, test & development set:* Since the available cross-linked data set is small in comparison with normal size data set to train deep learning models, a different approach will be used for this pipeline.

At first Text feature extraction part will be trained by TwitterText data set and next Picture feature extraction part will be trained by PsychoFlickr data set. When both parts have been trained the cross-linked Flickr-Twitter data set will be used for fine tuning the whole model with a low learning rate such as 0.01 or 0.001.

For training and test set for each neural network 5-fold cross-validation should be used.
Data sets should be divided as the table below

DataSet	Train set	Development set	Test set
TwitterText	60%	20%	20%
PsychoFlickr	60%	20%	20%
Flickr-Twitter	50%	25%	25%

Metrics

Evaluation metrics

Model and its components will be evaluated by Mean Absolute Error (MAE) and classification accuracy. Threshold for each component should be < 0.45

Deliverables

Step by step of the project process is

- Collecting Data sets. Since all the data sets are from research studies authors have been contacted to share their data set as of 01/29/2019
- Data preprocessing on datasets
- Data labeling on TwitterText data set
- Text neural network Implementation
- Training text neural network, evaluation, error analysis & tuning
- Picture neural network Implementation
- Training picture neural network, evaluation, error analysis & tuning
- Combination layer Implementation
- Model, evaluation & error analysis.

Error Analysis

There is number of ways we can do error analysis, one of the best ways is plotting cost function and the learning curve for different training sizes for training vs test set. These two plots will show whether or not the models are suffering from bias or variance and determine what direction should we move for performance improvement.

References

[1] Z. R. Samani, S. C. Guntuku, M. E. Moghaddam, D. Preoţiuc-Pietro, and L. H. Ungar, "Cross-platform and cross-interaction study of user personality based on images on twitter and flickr," *PloS one*, vol. 13, no. 7, p. e0198660, 2018.

[2] L. Liu, D. Preotiuc-Pietro, Z. R. Samani, M. E. Moghaddam, and L. H. Ungar, "Analyzing personality through social media profile picture choice." in *ICWSM*, 2016, pp. 211–220.

[3] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman *et al.*, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PloS one*, vol. 8, no. 9, p. e73791, 2013.

[4] D. Xue, L. Wu, Z. Hong, S. Guo, L. Gao, Z. Wu, X. Zhong, and J. Sun, "Deep learning-based personality recognition from text posts of online social networks," *Applied Intelligence*, pp. 1–15, 2018.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.