INF 385T Project Paper
Xiaoyue Pang, Xuechen Li

YouTube Vlogger’s Personality Prediction

Introduction

YouTube is the world’s largest video-sharing website, every day people watch hundreds of millions of hours on YouTube and generate billions of views. Most of the content on YouTube has been uploaded by individuals and many of them are vlogs that are performed by vloggers. The vlog platform is a booming market which allows people to express themselves freely. And the growth of vlogs bring us many interesting questions to study. Who are making vlogs? Why some of the top vloggers could earn such high volumes of views? Why people prefer a vlogger than another one?

We would like to make an initial attempt to solve these problems by analyzing the personalities of vloggers. We want to have an insight into the vlogger’s personality and hope it would help to explain some phenomenons. Moreover, we think that understanding the vloggers’ personalities could benefit further study and development about YouTube and other video-sharing websites by applying it to build recommendation system based on vloggers’ personalities. Other institutions could also use this method to learn people by the videos they made.

Overview

For this project, we will use both verbal data (manual transcripts) and nonverbal data (behavioral features) that could be extracted from a vlogger’s video as the source to predict one person’s personality. We will use labeled personality data so that we could apply supervised learning methods to build models.

This project will be consisted of 4 parts including preprocessing data, building nonverbal cues model, building verbal cues model, and combining models.

1. First, we will do a correlation analysis between nonverbal and gender features and Big-Five personality scores. The purpose is to find the mostsignificantly correlated features and leave out the ones that does not have correlation to any one of the personality scores. Then, we will divide each personality factor into three levels (high, medium and low) according to its score distribution.
2. Since the nonverbal dataset is non-lingrel, we are going to apply classification algorithms to build the models. Decision tree, k-NN, and SVM are algorithms we would like to try.
3. Then, we would like to tokenize raw text data, leave out the stop-words, vectorize the tokens, and extract features that are related to each personality factors. Then, we would try Naive Bayes model to classify texts into different levels of each personality factor.
4. Finally, based on the results of nonverbal cues model and verbal cues model, we would make a combination of the two parts to produce a better model for prediction.

Raw Dataset

The dataset is collected from 404 YouTube vloggers that explicitly show themselves in front of the a webcam talking about a variety of topics including personal issues, politics, movies, books, etc. There is no content-related restriction and the language used in the videos is natural and diverse. The dataset is divided into 4 parts: behavioral features, manual transcripts, personality impression scores, and gender.

1. Behavioral Features

Behavioral features described the nonverbal characteristics of a vlogger. The behavioral cues were automatically extracted from the conversational excerpts of vlogs and aggregated at video level.

a. Nonverbal features from audio

Audio cues were computed from the audio channel the toolbox developed by the Human Dynamics group at MIT Media Lab and include both speaking activity and prosody cues. This is a total of 21 features:

mean.pitch	sd.pitch	mean.conf.pitch	sd.conf.pitch	mean.spec.entropy	sd.spec.entropy
mean.val.apeak	sd.val.apeak	mean.loc.apeak	sd.loc.apeak	mean.num.apeak	sd.num.apeak
mean.energy	sd.energy	mean.d.energy	sd.d.energy	avg.voiced.seg	avg.len.seg
time.speaking	voice.rate	num.turns			

b. Nonverbal features from video

The overall motion of the vlogger is an indicator of vloggers' excitement and kinetic expressiveness. The overall visual activity of the vlogger was computed with a modified version of motion energy images called "Weighted Motion Energy Images" (wMEI). From the normalized wMEIs, statistical features were extracted as descriptors of the vlogger's body activity. This is a total of 4 features:

hogv.entropy	hogv.median	hogv.cogR	hogv.cogC

2. Manual Transcripts

A professional company manually transcribed the audio from vlogs. The transcription corresponds to the full video duration. In total, around 28h of video were annotated. Any name entities occurred in the video were anonymized by substituting with a 'XXXX'. The transcriptions are in raw text and contain a total of ~240K word tokens.

3. Personality Impression Scores

The personality impressions consist of Big-Five personality factors and each is valued by a score between 1 and 7. The scores were collected using Amazon Mechanical Turk and the Ten-Item Personality Inventory (TIPI). MTurk annotators watched one-minute slices of each vlog, and rated impressions using a personality questionnaire. The aggregated Big-Five scores are reliable with the following intra-class correlations (ICC(1,k), k=5): Extraversion (ICC = .76), Agreeableness (ICC = .64), Conscientiousness (ICC = .45), Emotional Stability (ICC = .42), Openness to Experience (ICC = .47), all significant with p < 10^{-3}.

4. Gender

The collection is mostly balanced in gender, with 194 males (48%%) and 210 females (52%). The gender labels are also provided.

Pre-processing

1. Unify the numeric format

The nonverbal data is consists of floating point numbers. For each column, there are several values have different decimal places. So the first pre-processing task is to unify the format of data in each column. We unified the decimal places by each column’s most common numeric format.

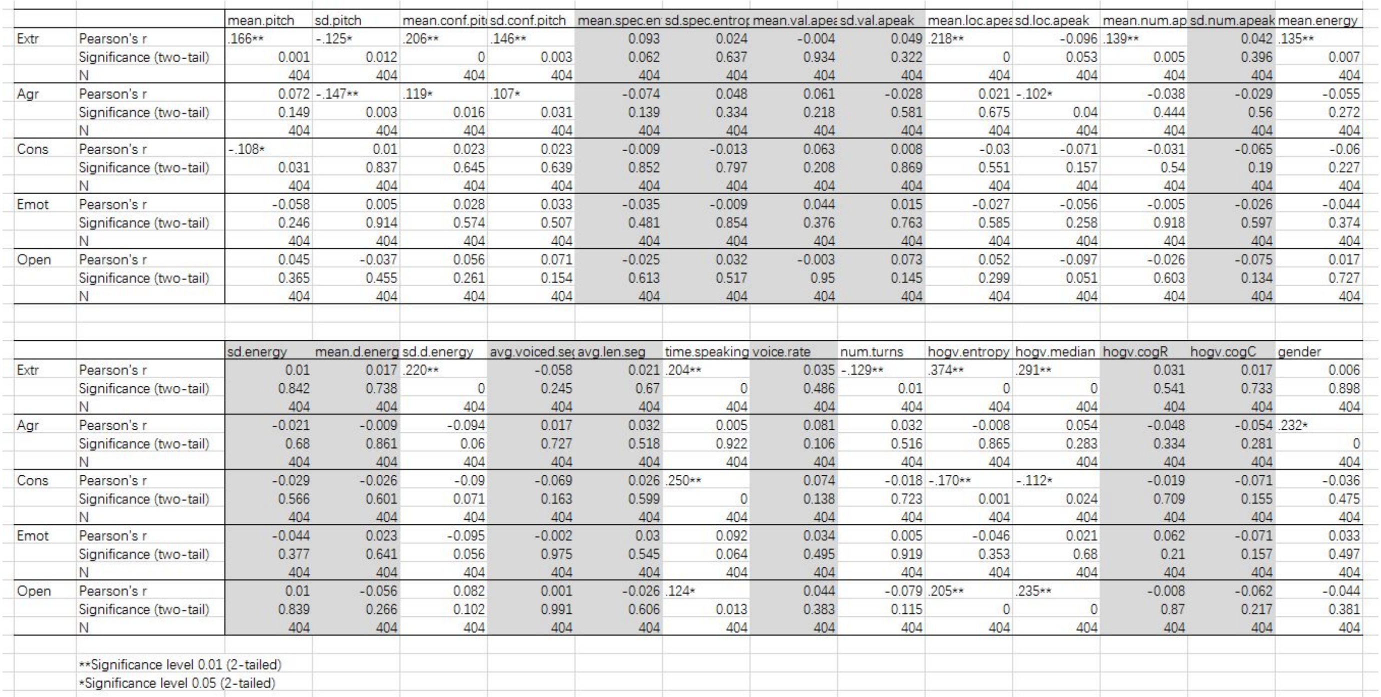
2. Join the nonverbal data and gender data

The gender data along is not helpful for predicting personality scores, so we think it is better to join gender data and nonverbal data together to build classification models. We coded gender data as 1 represent male and 2 represent female and added it to a column of the nonverbal data, treating the gender as a feature.

3. Perform correlation analysis

As part of our pre-processing process, we used SPSS to find the correlation among the features and the personality scores. We wanted to find the most correlated features and leave out the non-correlated ones in the hope to improve further built models’ performance. So we used the correlation analysis function in SPSS and it returned a form of Pearson product-moment correlation coefficient and Significance values (two-tail), as the figure 1 shown below. We found that the following features does not have correlation to any one of the personality scores. So we removed these columns from our nonverbal data.

mean.spec.entropy	sd.spec.entropy	mean.val.apeak	sd.val.apeak
sd.num.apeak	sd.energy	mean.d.energy	avg.voiced.seg
avg.len.seg	voice.rate	hogv.cogR	hogv.cogC



		mean.pitch	sd.pitch	mean.conf.pit	sd.conf.pitch	mean.spec.en	sd.spec.entro	mean.val.apea	sd.val.apeak	mean.loc.apea	sd.loc.apeak	mean.num.ap	sd.num.apeak	mean.energy
Extr	Pearson's r	.166**	-.125*	.206**	.146**	0.093	0.024	-0.004	0.049	.218**	-0.096	.139**	0.042	.135**
	Significance (two-tail)	0.001	0.012	0	0.003	0.062	0.637	0.934	0.322	0	0.053	0.005	0.396	0.007
	N	404	404	404	404	404	404	404	404	404	404	404	404	404
Agr	Pearson's r	0.072	-.147**	.119*	.107*	-0.074	0.048	0.061	-0.028	0.021	-.102*	-0.038	-0.029	-0.055
	Significance (two-tail)	0.149	0.003	0.016	0.031	0.139	0.334	0.218	0.581	0.675	0.04	0.444	0.56	0.272
	N	404	404	404	404	404	404	404	404	404	404	404	404	404
Cons	Pearson's r	-.108*	0.01	0.023	0.023	-0.009	-0.013	0.063	0.008	-0.03	-0.071	-0.031	-0.065	-0.06
	Significance (two-tail)	0.031	0.837	0.645	0.639	0.852	0.797	0.208	0.869	0.551	0.157	0.54	0.19	0.227
	N	404	404	404	404	404	404	404	404	404	404	404	404	404
Emot	Pearson's r	-0.058	0.005	0.028	0.033	-0.035	-0.009	0.044	0.015	-0.027	-0.056	-0.005	-0.026	-0.044
	Significance (two-tail)	0.246	0.914	0.574	0.507	0.481	0.854	0.376	0.763	0.585	0.258	0.918	0.597	0.374
	N	404	404	404	404	404	404	404	404	404	404	404	404	404
Open	Pearson's r	0.045	-0.037	0.056	0.071	-0.025	0.032	-0.003	0.073	0.052	-0.097	-0.026	-0.075	0.017
	Significance (two-tail)	0.365	0.455	0.261	0.154	0.613	0.517	0.95	0.145	0.299	0.051	0.603	0.134	0.727
	N	404	404	404	404	404	404	404	404	404	404	404	404	404

		sd.energy	mean.d.energ	sd.d.energy	avg.voiced.se	avg.len.seg	time.speaking	voice.rate	num.turns	hogv.entropy	hogv.median	hogv.cogR	hogv.cogC	gender
Extr	Pearson's r	0.01	0.017	.220**	-0.058	0.021	.204**	0.035	-.129**	.374**	.291**	0.031	0.017	0.006
	Significance (two-tail)	0.842	0.738	0	0.245	0.67	0	0.486	0.01	0	0	0.541	0.733	0.898
	N	404	404	404	404	404	404	404	404	404	404	404	404	404
Agr	Pearson's r	-0.021	-0.009	-0.094	0.017	0.032	0.005	0.081	0.032	-0.008	0.054	-0.048	-0.054	.232*
	Significance (two-tail)	0.68	0.861	0.06	0.727	0.518	0.922	0.106	0.516	0.865	0.283	0.334	0.281	0
	N	404	404	404	404	404	404	404	404	404	404	404	404	404
Cons	Pearson's r	-0.029	-0.026	-0.09	-0.069	0.026	.250**	0.074	-0.018	-.170**	-.112*	-0.019	-0.071	-0.036
	Significance (two-tail)	0.566	0.601	0.071	0.163	0.599	0	0.138	0.723	0.001	0.024	0.709	0.155	0.475
	N	404	404	404	404	404	404	404	404	404	404	404	404	404
Emot	Pearson's r	-0.044	0.023	-0.095	-0.002	0.03	0.092	0.034	0.005	-0.046	0.021	0.062	-0.071	0.033
	Significance (two-tail)	0.377	0.641	0.056	0.975	0.545	0.064	0.495	0.919	0.353	0.68	0.21	0.157	0.497
	N	404	404	404	404	404	404	404	404	404	404	404	404	404
Open	Pearson's r	0.01	-0.056	0.082	0.001	-0.026	.124*	0.044	-0.079	.205**	.235**	-0.008	-0.062	-0.044
	Significance (two-tail)	0.839	0.266	0.102	0.991	0.606	0.013	0.383	0.115	0	0	0.87	0.217	0.381
	N	404	404	404	404	404	404	404	404	404	404	404	404	404

**Significance level 0.01 (2-tailed)
*Significance level 0.05 (2-tailed)

Figure 1. The correlation results of each feature

We also used Python to do the correlation analysis by using the scipy.stats.pearsonr function in SciPy. We added the gender data as the last column of the features, which means the raw features are from column 1 to column 26. And the result shows that columns [1, 2, 3, 4, 9, 10, 11, 13, 16, 19, 21, 22, 23, 26] have 2-tailed p-value <= 0.05 . Now we have 14 features that can be considered correlated to the personality scores. It is same as the result from SPSS.

4. Transfer verbal data into a list

Our verbal data was initially stored in 404 txt file separately. We read and transferred the data into a list. It helped us to do further vectorizing task.

Initial Analyses

We wanted to see vloggers' personality data structure, in order to decide how to divide them into different categories. We found that most scores distributed in the normal distribution way, as shown in figure 2.

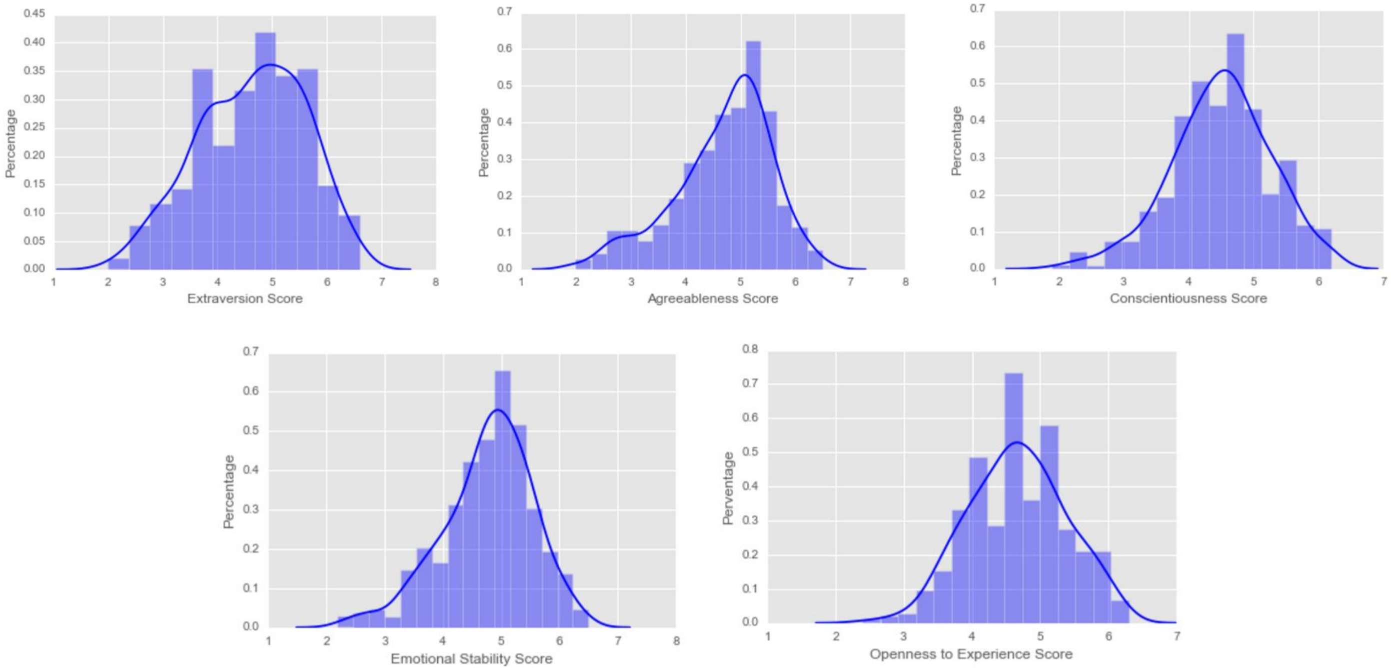


Figure 2. The individual percentage histogram with curve of each score

We set vloggers' personality scores into three levels as the labels for the classification models. In order to set each level evenly, we plotted the cumulative percentage of each score, shown in figure 3. We labeled the lowest one third of the scores as the low level, the highest one third as the high level, and then the middle part as the medium level.

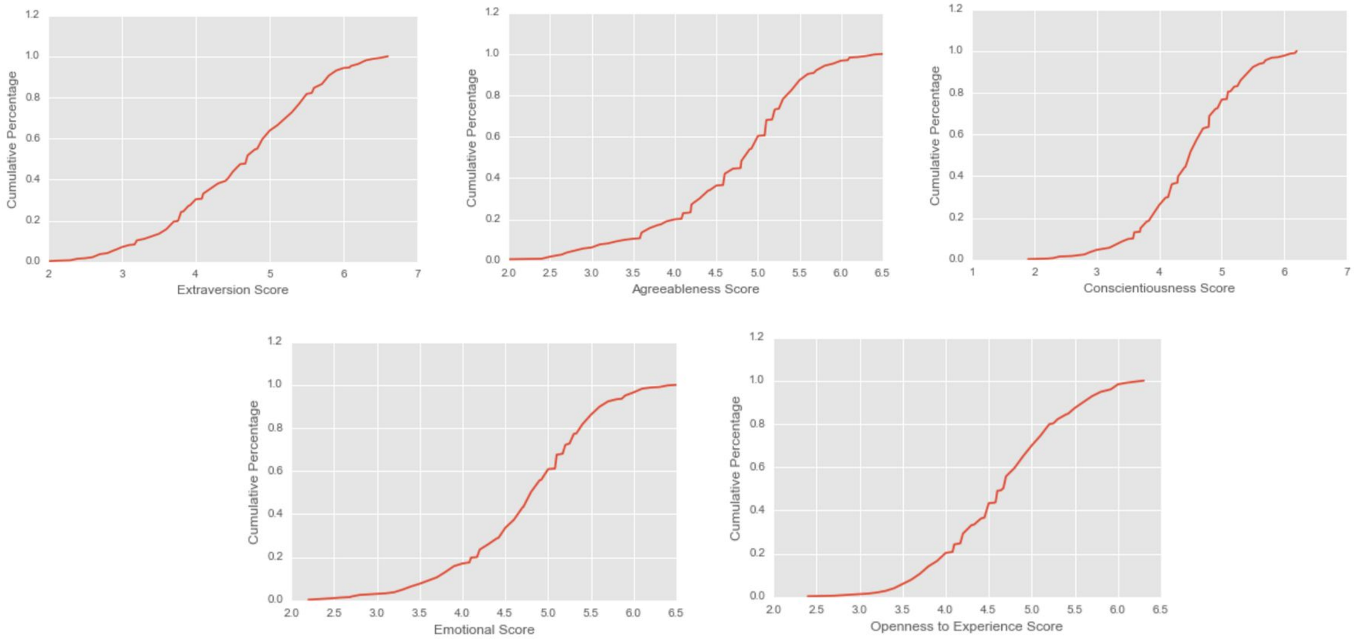


Figure 3. The cumulative percentage curve of each score

To see the distribution of vloggers' personality levels related to their nonverbal features in the video, we first reduced nonverbal data into 2 dimensions using PCA and plotted it out, marked dots with different colors regarding to the personality levels.

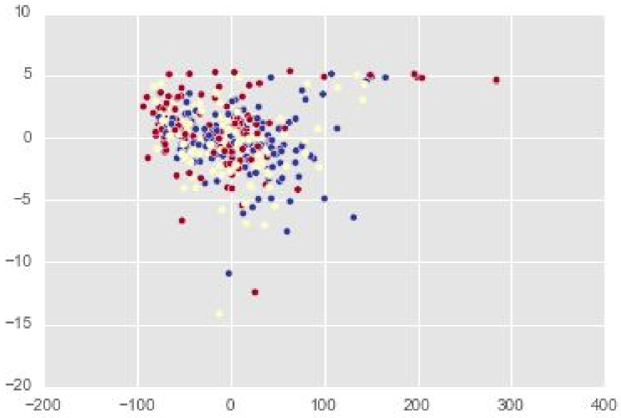


Figure 4. 2-dimension nonverbal data, color marked by their personality levels

As shown in figure 4, dots in three different levels are not so well separated to cluster into obvious clusters. It probably because the data has many dimensions and the labeled data is continuous.

Preliminary Results

We tried to build models separately on nonverbal with gender data and verbal data.

1. Nonverbal with gender data

First, we split our nonverbal with gender data into two parts, 0.7 of the dataset is treated as train data and the rest 0.3 of the dataset is the test data.

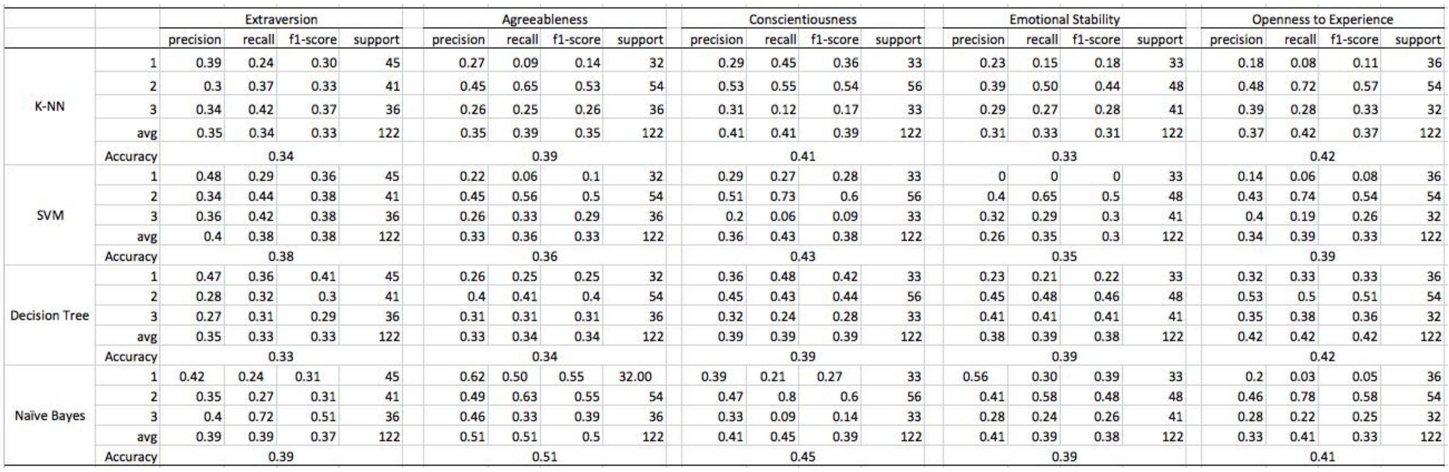
Second, we applied several different classification algorithms, K-NN, SVM and Decision tree to build classification models. We applied cross-fold validation of 10 to each one of the model. For each method, we built five models to predict five personality scores, which are Extraversion, Agreeableness, Conscientiousness, Emotional stability, Openness to experience. For the K-NN method, we performed calculations to find the best n value in the range from 1 to 15 for each personality score and applied it in the model building. Then we applied evaluation methods to see the models' performance.

2. Verbal data

We used Naive Bayes method to train verbal model. After vectorizing all the words in transcripts, we fitted the Naive Bayes models on our training dataset. Then we tested the trained models on test dataset, and found their performance did a little better job than nonverbal models (shown in Figure 5).

3. Evaluation methods

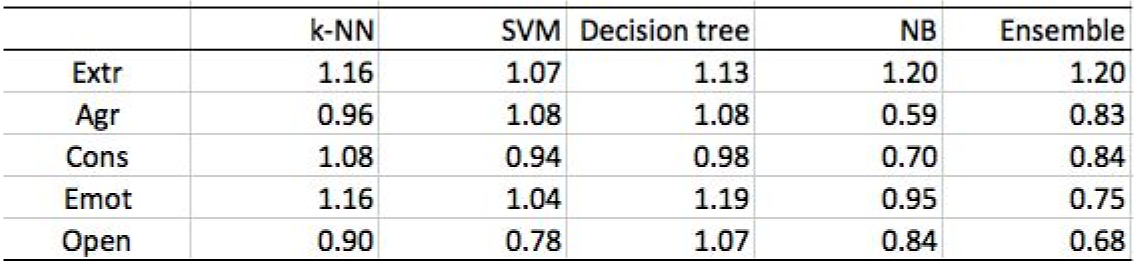
At first, we chose to use accuracy and confusion metrics to measure each model's performance. We found that most of our models could do the classification better than chance, but not too much. The accuracy and precision of each model seems to be similar. There is no obvious evidence that shows which one of the models is better than the others. Concretely. Figure 5 shows the performance metrics of K-NN, SVM and DT models of each personality score.



		Extraversion				Agreeableness				Conscientiousness				Emotional Stability				Openness to Experience			
		precision	recall	f1-score	support	precision	recall	f1-score	support	precision	recall	f1-score	support	precision	recall	f1-score	support	precision	recall	f1-score	support
K-NN	1	0.39	0.24	0.30	45	0.27	0.09	0.14	32	0.29	0.45	0.36	33	0.23	0.15	0.18	33	0.18	0.08	0.11	36
	2	0.3	0.37	0.33	41	0.45	0.65	0.53	54	0.53	0.55	0.54	56	0.39	0.50	0.44	48	0.48	0.72	0.57	54
	3	0.34	0.42	0.37	36	0.26	0.25	0.26	36	0.31	0.12	0.17	33	0.29	0.27	0.28	41	0.39	0.28	0.33	32
	avg	0.35	0.34	0.33	122	0.35	0.39	0.35	122	0.41	0.41	0.39	122	0.31	0.33	0.31	122	0.37	0.42	0.37	122
	Accuracy	0.34				0.39				0.41				0.33				0.42			
SVM	1	0.48	0.29	0.36	45	0.22	0.06	0.1	32	0.29	0.27	0.28	33	0	0	0	33	0.14	0.06	0.08	36
	2	0.34	0.44	0.38	41	0.45	0.56	0.5	54	0.51	0.73	0.6	56	0.4	0.65	0.5	48	0.43	0.74	0.54	54
	3	0.36	0.42	0.38	36	0.26	0.33	0.29	36	0.2	0.06	0.09	33	0.32	0.29	0.3	41	0.4	0.19	0.26	32
	avg	0.4	0.38	0.38	122	0.33	0.36	0.33	122	0.36	0.43	0.38	122	0.26	0.35	0.3	122	0.34	0.39	0.33	122
	Accuracy	0.38				0.36				0.43				0.35				0.39			
Decision Tree	1	0.47	0.36	0.41	45	0.26	0.25	0.25	32	0.36	0.48	0.42	33	0.23	0.21	0.22	33	0.32	0.33	0.33	36
	2	0.28	0.32	0.3	41	0.4	0.41	0.4	54	0.45	0.43	0.44	56	0.45	0.48	0.46	48	0.53	0.5	0.51	54
	3	0.27	0.31	0.29	36	0.31	0.31	0.31	36	0.32	0.24	0.28	33	0.41	0.41	0.41	41	0.35	0.38	0.36	32
	avg	0.35	0.33	0.33	122	0.33	0.34	0.34	122	0.39	0.39	0.39	122	0.38	0.39	0.38	122	0.42	0.42	0.42	122
	Accuracy	0.33				0.34				0.39				0.39				0.42			
Naïve Bayes	1	0.42	0.24	0.31	45	0.62	0.50	0.55	32.00	0.39	0.21	0.27	33	0.56	0.30	0.39	33	0.2	0.03	0.05	36
	2	0.35	0.27	0.31	41	0.49	0.63	0.55	54	0.47	0.8	0.6	56	0.41	0.58	0.48	48	0.46	0.78	0.58	54
	3	0.4	0.72	0.51	36	0.46	0.33	0.39	36	0.33	0.09	0.14	33	0.28	0.24	0.26	41	0.28	0.22	0.25	32
	avg	0.39	0.39	0.37	122	0.51	0.51	0.5	122	0.41	0.45	0.39	122	0.41	0.39	0.38	122	0.33	0.41	0.33	122
	Accuracy	0.39				0.51				0.45				0.39				0.41			

Figure 5. Evaluation of each model's performance

Then we realized accuracy and confusion metrics are not everything. Accuracy does not make sense especially in terms of three categories classification task. Both the conditions where the model predicts outcomes as high and medium when the true value is low are considered as the same type of fault by accuracy. We cannot know the difference between prediction and true value from accuracy and confusion metrics. So we tried to use mean-squared error (MSE) to measure models' performance. As shown in Figure 6, mean-squared errors of k-NN, SVM, Decision tree and Naive Bayes are around 1 and have slight difference across each model. Overall, SVM does better job on nonverbal and gender data than other two models; Naive Bayes model trained on verbal data has better prediction of the outcome compared to nonverbal and gender model.



	k-NN	SVM	Decision tree	NB	Ensemble
Extr	1.16	1.07	1.13	1.20	1.20
Agr	0.96	1.08	1.08	0.59	0.83
Cons	1.08	0.94	0.98	0.70	0.84
Emot	1.16	1.04	1.19	0.95	0.75
Open	0.90	0.78	1.07	0.84	0.68

Figure 6. MSE of each model

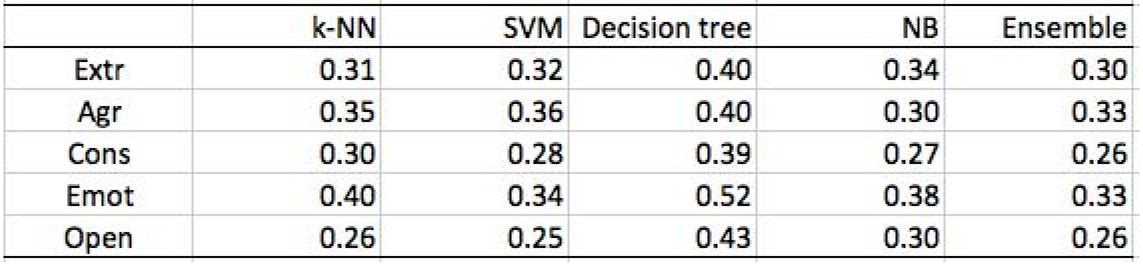
4. Models combination

To complete our model, we wanted to combine the information two separate models together. We used hstack from scipy to link two vectors from nonverbal and gender data and verbal data to be a new vector as the input. Then we applied voting classifier to combine k-NN, SVM, DT and NB models as an ensemble model. We set the weight of each model as 1, since there is no significant difference between each model’s performance. We trained the ensemble model on the new linked training dataset and tested it on the new linked testing dataset. To measure the ensemble models’ performance, we calculated the MSE of each personality score’s model (shown in Figure 6). Except that the prediction of Extraversion score is a little poorer than others, the mean-squared errors are almost less than 1.

Improvement

Given the size of our data, we may not be able to train predictive enough models on vloggers’ personality in this case. So we would like to lower the difficulty of our task. To gain a better performance of our model, we narrowed down the number of personality levels from three to two. We kept the high level and combine the medium and low levels into “not high” category. Then our task is to predict whether the vloggers have high scores in the five aspects.

We applied the same methods on the binary dataset and got better performance on each model in terms of MSE. From the Figure 7, we found k-NN and SVM performed better than Decision tree model on nonverbal and gender data. Naive Bayes model on verbal data performed almost the same as k-NN and SVM. We equally combined these four models together and got a new binary ensemble model. As we expected, the model performed better in binary classification task. Even though we cut down the number of personality levels, the result is still meaningful. As people with high scores in some personality aspect may show uncommon characteristics, the prediction of this crowd of people could be informative.



	k-NN	SVM	Decision tree	NB	Ensemble
Extr	0.31	0.32	0.40	0.34	0.30
Agr	0.35	0.36	0.40	0.30	0.33
Cons	0.30	0.28	0.39	0.27	0.26
Emot	0.40	0.34	0.52	0.38	0.33
Open	0.26	0.25	0.43	0.30	0.26

Figure 7. MSE of each binary model

Discussion

We have given a lot of thoughts to this project. First we tried to decide whether to use classification or regression methods to make prediction. Given the features and characteristics of our data, we decided to make it a classification work. Then we made efforts to find a good way to categorize the labeled data without losing its characteristics and making the task fair enough to achieve our goal. We also have been thinking about how to combine the nonverbal and verbal parts, how to ensemble the different models to produce a better model. We believe that the most important thing about our project is that we proved that the observing record data of a person could be meaningful as much as a self-reported metrics to reflect one’s personality. It is interesting to see that information that seems as subjective as a personality trait can be analyzed by machine learning.

We think that the method we are using to describe a vlog is well established. By using these features, a vlogger’s behavior, voice, and the script he/she said can be extracted. It is also a good practice to use Amazon’s Mechanical Turk to label the personality traits with scores. People’s personality is a complex topic, and the short duration of a video make it harder to extract enough data. So the next step should be first collecting more data to form a larger sample. A considerable volume of data of more variance would make it possible to perform other machine learning methods like regression methods. We believe that further study on this topic could also try to extract expression features from the vloggers, and their verbal data could be further explored by using some sentiment analyzing methods.