


# Predicting well-being from Instagram data

Miltenburg Lino<sup>1[11136375]</sup>, Tsimpoukis Dimitrios<sup>1[12338281]</sup>, Schniepp Michael<sup>1[12160067]</sup>, Symeonidou Anthi<sup>1[12296082]</sup>, and Wilmers Ben<sup>1[12292117]</sup>

University of Amsterdam, Amsterdam, NL

**Abstract.** Photo and video sharing on social media platforms such as Instagram has become a very common practice worldwide and as a result has become a very important study source for social analytics. In this work, we assess the ability to predict the well-being of a given Instagram user based on their uploaded image content. Different datasets focusing on different image material were utilized, resulting in two predictive models. Even though the results were not ideal, we gained fruitful insight into how different data from social media can be interpreted, utilized and manipulated to formulate prediction models.

**Keywords:** Instagram data · well-being · regression model · machine-learning · survey

The GitHub  repository containing the source code for this study can be found in this [link](#)

## 1 Introduction

Instagram is one of the most used social media platforms worldwide, where users share photos and videos. A large amount of visual image data is generated and can be used for research. A recent study has shown a correlation between Instagram posts and clinical markers of depression [4]. This paper assesses the correlation between the Instagram activity of a given user, the content of their images and their psychological well-being. Well-being is quantified by means of a score (*PERMA*) from the results of a survey on 159 Instagram users. The goal of this study is to evaluate the relationship between the happiness or well-being of the Instagram user and their image information and content.

## 2 Methodology

For this study, the well-being of the 159 persons that completed the PERMA survey was evaluated. Six different datasets were used, each containing information pertaining to the survey respondent's Instagram posts. First of all, each dataset required cleaning and preprocessing for further analysis. Each dataframe underwent its own preprocessing detailed in the sections below. Eventually, the datasets had to be aggregated to user level in order to find a connection with

an individual's photos and their numerical well-being scores. After preprocessing, a selection of variables was made and the datasets were merged into a final dataframe to construct a model and assess the predictive ability of the Instagram data.

## 2.1 Original dataframes' preprocessing

**Survey** Well-being is measured through a survey using the PERMA scale. Users from crowd-sourcing platforms (Mechanical Turk and Microwork) were asked to log in with their Instagram account and fill in the PERMA survey. The PERMA scale was developed by Martin Seligman and takes into account five core-elements of psychological well-being and happiness [6]. These are positive and negative emotions, engagement, relationships, meaning, accomplishment and health. The PERMA score is the dependent variable in this assessment and has been calculated as the mean of all the PERMA subscores.

**Image metrics** The image metrics data set contains information about the amount of likes and comments on the images at various points in time. Only the most recent like and comment count for each picture was considered. These were used to calculate mean values for the images posted by each user.

**Image** The image dataset contains metadata of the images such as the url, the dimensions and the date and time of an image. Many of these variables, such as date or url, are not usable to construct a predictive model and are therefore dropped from the dataframe. The correlation matrix showed a negative correlation between `user_posted_photos` and the PERMA-score. The parameters `user_followed_by` and the `user_followers` showed a lower correlation but were still selected for the final dataframe after each variable was aggregated to the user level.

**Face** The face dataset contains information about the face detection that has been executed on each image. The dataframe contains 86877 rows with 15957 identical images. The images can contain multiple faces, and for each face the top three of the face emotion has been determined with a confidence score. The face emotion with the highest confidence score has been aggregated and eventually these categorical variables have been turned into an indicator variable. Other detection variables in the dataset are: `beard`, `mustache`, `smile`, `eyeglasses`, `sunglasses`, `gender` and `age`. All image detection variables are denoted with a true/false indicator and an accompanying confidence score. These confidence scores were dropped as they are not explanatory. The sum of all indicator variables per user were eventually aggregated to a percentage of images containing these variables per user. For example, the percentage of images that contained eyeglasses per user. In the end, the final face dataframe contained information about 145 unique users.

**Object Labels** The Object Labels dataframe contains information regarding the content of the instagram images. Unknown image recognition software was used to classify certain objects that may be present in the image. For each image up to 10 objects were identified with at least 70% confidence. Thus each image id is paired with an object label and confidence level.

The objective here was to distill the over 42,000-image dataframe down into user-level detail that could possibly offer some insight into the overall condition of the user. The unaltered dataframe contains over 2000 unique object labels. Upon inspection of the labels ordered by frequency of use, we found that most labels were used very rarely and a small number of labels used very often, with frequencies dropping exponentially from the top of the list. Because of this, we did not think we could gain much information from labels that appear a handful of times (most only once) out of over 42,000 rows. From here we decided to take the top 50 most used labels and subsequently one hot encode them. This resulted in marking each image with a binary variable indicating the presence of one of these labels. Once the labels were encoded we could take an aggregate of label counts by user. Next, we standardized these label frequencies by dividing by the total number of posts that a user has resulting in a ratio describing the percent of photos in which said label appears for each user. Using this method we could then determine the general content most prevalent in a users posts based on the image recognition classifications in a format friendly for statistical analysis and machine learning algorithms.

**Adjective Noun Pairs (ANP)** The ANP dataset includes different adjective-noun pairs associated with every image. Each ANP is given an emotion label and an emotion score based on Plutchik’s defined 24 emotions [3]. The different emotions are depicted in the Wheel of Emotions below.

In order to be able to quantify better the different emotion labels, we decided to cluster them into 4 different categories and assign a score into each label’s value based on the emotion they are describing. The 4 new categories we concluded upon were Aggressiveness, Attention, Pleasantness and Appreciation. The different labels associated with each cluster are presented in Table 1.

The 4 new categories are added as new columns in the ANP dataframe and correspond to 4 new features that we used in our model. After correlation testing the `anp_label` column was dropped due to high correlation to the 4 new features as expected. Finally, each of the remaining columns was aggregated based on the `user_id` using the mean values. In the case of the 4 new features (**Aggressiveness**, **Attention**, **Pleasantness** and **Appreciation**) the resulting NaN values of the label clustering procedure are replaced by the value 0.



Fig. 1: Plutchnik's Wheel of Emotions

Class / Score	-3	-2	-1	1	2	3
Aggressiveness	terror	fear	apprehension	annoyance	anger	rage
Attention	amazement	surprise	distraction	interest	anticipation	vigilance
Pleasantness	grief	sadness	pensiveness	serenity	joy	ecstasy
Appreciation	loathing	disgust	boredom	acceptance	trust	admiration

Table 1: New sentiment label clusters and associated value scores

## 2.2 Merging of dataframes and model selection

Following the preprocessing of all the separate dataframes they were all merged into one final big dataframe consisting of 99 features. The final dataframe required some further preprocessing before creating the regression models.

**Feature Correlation check** The first thing that we needed to ensure after the merging was that there were no strong correlations between features. This became essential due to the encoding of various categorical variables that resulted in immediate correlations (e.g `face_beard` and `face_mustache`). Applying the step-up method between all the highly correlated tuples, we checked their separate significance towards the overall PERMA score and we eventually dropped

the least significant of the two features of each tuple with correlation higher than 0.4 between each other. Our final feature pool consisted of 43 features.

**Final Feature Selection** For our final model feature selection, we initially followed a typical step-down approach in which case we started with the entire final feature pool and eliminated step-by step the features based on the  $P - values$  towards the PERMA score [8]. The model that was generated from the result of this approach was not satisfactory, yielding predictions on the final PERMA scores with *Root-Mean-Square-Error (RMSE)* above 1.7 so we decided to follow a different approach.

We decided to use one of the feature selection wrappers from `scikit-learn` named `SelectKBest` [2]. This wrapper performs automatic feature selection given an input of features and target outputs, and tries to find dependencies between the features and their impact towards the desired output. The selection is based on a particular metric such as *ANOVA F-values* among others. After testing the selections based on different metrics we concluded in the mutual information (MI) metric. It is a non-parametric method based on entropy estimation from k-nearest neighbors distances and performed well in our case due to the fact that even after aggregation many of the features that were the result of encoding still had categorical variable characteristics such as discrete values [5].

**Linear-Regression Model** By using `SelectKBest` and a grid-search algorithm we generated various linear regression models based on the selected number of features generated by `K-Best`. For each model the corresponding features and outputs dataset were split into a training and test set with a ratio of 0.8. Our final model selection choice was based on a tradeoff between the Root Mean Square Error, Mean Absolute Error between prediction and real values of the test set, and the explained variation (Adjusted  $R^2$ ) [8,7]. The final selected model consisted of 18 features which are shown in Table 2.

**Support Vector Regression Model** On top of our linear regression model we decided to also incorporate a Support Vector Regression (SVR) model in our approach. Support Vector Machine methods utilize hyperplanes and are capable of performing both classification as well as regression tasks [1]. In our example we tested 3 different kernels (linear, polynomial and Radial basis function) for our SVR model and tuned the hyperparameters with a grid search algorithm. We concluded on an SVR model with Radial basis function (rbf) kernel and hyperparameter value  $C = 1000$ .

**Cross-Validation** The result metrics from our model after training give a numerical estimate of the difference between predicted and original values based on the whole set of training data. However, this is an indication of the performance of the model trained on one particular set of data (training set) and the danger

	Feature
0	user_followed_by
1	data_amz_label_Couch
2	Aggressiveness
3	Pleasantness
4	Appreciation
5	face_sunglasses_perc
6	income_\$90,000 to \$99,999
7	face_emo_DISGUSTED_perc
8	employed_Out of work but not currently looking...
9	data_amz_label_Bottle
10	education_High school graduate
11	data_amz_label_Clothing
12	employed_Employed for wages
13	employed_Out of work and looking for work
14	employed_A student
15	income_\$100,000 to \$149,999
16	education_Post graduate degree
17	user_posted_photos

Table 2: Final Feature Selection

of under/overfitting is always present. Especially in our case, where the dataset is small (159 users) the need for further validation that determines whether the model will be able to generalize to unseen data is more evident. To that regard, we will *Cross-Validate* our model utilizing the K-Folds CV technique. According to this technique, the original *training set* is split into  $k$  training sets and corresponding test/validation sets and the generated model is fitted  $k$  times. The error metric of each model is averaged between the  $k$  trials and if its value is close to the one suggested from the actual prediction results against the original test-set, then the model is validated. In our case we used 5 folds validation and the error metric for cross validation was the Mean Absolute Error.

### 3 Results

#### 3.1 Linear Regression

The accuracy results that determine the performance of our linear regression model are shown in Table 3. Also the regression coefficients as well as some feature specific statistical information are presented in Figure 2.

Prediction RMSE	Prediction Mean Abs. Error	$R^2$	Adjusted $R^2$	Cross Validation Mean Abs. Error
1.417899	1.060231	0.252967	0.128461	1.072778

Table 3: Linear Regression Model Scores

	coef	std err	t	P> t	[0.025	0.975]
const	7.2023	0.416	17.297	0.000	6.377	8.028
user_followed_by	2.234e-05	0.000	0.106	0.916	-0.000	0.000
data_amz_label_Couch	-3.8488	2.973	-1.295	0.198	-9.742	2.044
Aggressiveness	0.1619	0.140	1.154	0.251	-0.116	0.440
Pleasantness	0.4472	0.291	1.536	0.128	-0.130	1.024
Appreciation	-0.2088	0.195	-1.068	0.288	-0.596	0.179
face_sunglasses_perc	-0.0698	1.517	-0.046	0.963	-3.078	2.938
income_90,000to99,999	0.5149	0.633	0.814	0.417	-0.739	1.769
face_emo_DISGUSTED_perc	-0.6447	8.770	-0.074	0.942	-18.028	16.739
employed_Out of work but not currently looking for work	-0.0415	0.737	-0.056	0.955	-1.502	1.419
data_amz_label_Bottle	-4.5364	6.471	-0.701	0.485	-17.364	8.291
education_High school graduate	-0.5234	0.306	-1.710	0.090	-1.130	0.083
data_amz_label_Clothing	-2.8573	1.429	-1.999	0.048	-5.691	-0.024
employed_Employed for wages	0.0055	0.300	0.018	0.985	-0.589	0.600
employed_Out of work and looking for work	-0.9095	0.609	-1.494	0.138	-2.116	0.297
employed_A student	-0.3742	0.507	-0.739	0.462	-1.378	0.630
income_100,000to149,999	0.7800	0.421	1.851	0.067	-0.055	1.615
education_Post graduate degree	0.4787	0.507	0.944	0.347	-0.526	1.484
user_posted_photos	-0.0006	0.000	-1.919	0.058	-0.001	1.93e-05

Fig. 2: Statistical Summary of the Linear Regression Model's features

Based on the Adjusted  $R^2$  value, we can see that the model explains around 13% of the variation in the PERMA scores of our data. This is far from ideal, but the best we could achieve in our case. Given the small amount of data and the large number of features it is rather difficult to achieve high levels of  $R^2$ . Also, since prediction model accuracy was the most desired attribute of our model we focused more on minimizing the RMSE and MAE in our solution. The RMSE error value of 1.417899 is relatively higher than the 1.060231 value of the MAE but this is expected since the RMSE punishes higher error values. In our case, since the PERMA score is measured on a scale from 1-10 we can conclude that a MAE value of 1.06 is not very good but shows a decent predictive capability from our model. In terms of validation, the cross-validation averaged MAE we can see that the average MAE value is very close to the one we obtained from testing our model. This validates our model and makes it applicable for predictions on unseen data.

### 3.2 Support Vector Regression (SVR)

The results for the Support Vector Regression are shown in Table 4. We observe a slight improvement in terms of RMSE and MAE. What is interesting though is that we observe a small but more distinct discrepancy between the cross-validation averaged MAE and the prediction MAE. This suggests a small overfit from SVR. However, given our very small amount of data this is more than expected.

Prediction RMSE	Prediction Mean Abs. Error	Cross Validation Mean Abs. Error
1.321152	1.013761	1.150853

Table 4: Support Vector Regression Model Scores

## 4 Discussion

Upon cleaning and analysing the given datasets we discovered that the datasets individually showed very little relationship to the PERMA score or it’s constituent P, E, R, M, and A scores. Upon merging the dataframes and further preprocessing we concluded on a model of 18 features. The prediction results in terms of explained variation were far from satisfactory, however our models showed some predictive ability.

There are a number of factors that hindered formulating a more robust relationship between features and PERMA scores. First of all, there will always be an inherent loss of information when many rows of data must be aggregated, in order to profile a single individual. Moreover, we believe that the PERMA scores themselves are not strong or realistic indicators of a person’s well-being, given that the responses are self-reported which indicates an inherent level of bias or inaccuracy. In addition, question responses can be open to interpretation and translating them into a numerical value can further result in loss or distortion of information. Finally, the resulting dataframe of the most relevant features could be considered reasonably small at fewer than 160 rows.

Taking into account the above considerations, the necessity of further *Feature Engineering* becomes evident. Given that due to time limitations we could not expand into further feature extraction from the given datasets, we will present some thoughts on how the model could be improved. First of all, there should be some association of the image data based on the date the image was posted related to the date of the survey. For instance, image data should be weighted more if the photo was posted in dates relatively close to the survey. A person’s well being/mood can change dramatically between time periods. One other possible feature that could be extracted would be the appearances of the user himself in his uploaded pictures. Given that there is a **face\_id** attribute in the **face** dataframe, we could assume that the face\_id with the most appearances for a particular **user\_id** belongs the user himself.

To conclude, from our analysis, even though we were not able to pinpoint a very accurate relationship between the image meta-data and the respondent’s general well-being, we were able to gain some important insights from it. Our feature selection process showed relations between user well-being and factors such as education, employment status, number of followers and images posted. As a final thought, we should say that even though social network analysis can yield very interesting results related to topics such as social dynamics, a particular user’s well being is a relatively abstract and complex topic for accurate analysis.



However, with large datasets, much more sophisticated prediction models and proper feature engineering more accurate results could be achieved.

## References

1. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg (2006)
2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)
3. Plutchik, R.: A psychoevolutionary theory of emotions. *Social Science Information* **21**(4-5), 529–553 (1982). <https://doi.org/10.1177/053901882021004003>, <https://doi.org/10.1177/053901882021004003>
4. Reece, A.G., Danforth, C.M.: Instagram photos reveal predictive markers of depression. *CoRR* **abs/1608.03282** (2016), <http://arxiv.org/abs/1608.03282>
5. Ross, B.C.: Mutual information between discrete and continuous data sets. *PLOS ONE* **9**(2), 1–5 (02 2014). <https://doi.org/10.1371/journal.pone.0087357>, <https://doi.org/10.1371/journal.pone.0087357>
6. Seligman, M.: Flourish : a visionary new understanding of happiness and well-being. Free Press, New York (2011)
7. Triola, M.: Elementary Statistics. MyStatLab Series, Pearson/Addison-Wesley (2006), [https://books.google.nl/books?id=YE\\_sAQAAIAAJ](https://books.google.nl/books?id=YE_sAQAAIAAJ)
8. Vik, P.: Regression, ANOVA, and the General Linear Model: A Statistics Primer. SAGE Publications (2013), <https://books.google.nl/books?id=Lox3ngEACAAJ>