Are You Satisfied with Life?: Predicting Satisfaction with Life from Facebook

Susan Collins[1,2], Yizhou Sun[1], Michal Kosinski[3], David Stillwell[3], Natasha Markuzon[2⋆]

[1] Northeastern University, CCIS, Boston, MA, USA
{skcoll,yzsun}@ccs.neu.edu
[2] Draper Laboratory, Boston, MA, USA
{skcollins,nmarkuzon}@draper.com
[3] Free School Lane, The Psychometrics Centre, University of Cambridge, Cambridge CB2 3RQ United Kingdom
{mk583,ds617}@cam.ac.uk

Abstract. Social media can be beneficial in detecting early signs of emotional difficulty. We utilized the Satisfaction with Life (SWL) index as a cognitive health measure and presented models to predict an individual's SWL. Our models considered ego, temporal, and link Facebook features collected through the myPersonality.org project. We demonstrated the strong correlation between Big 5 personality features and SWL, and we used this insight to build two-step Random Forest Regression models from ego features. As an intermediate step, the two-step model predicts Big 5 features that are later incorporated in the SWL prediction models. We showed that the two-step approach more accurately predicted SWL than one-step models. By incorporating temporal features we demonstrated that "mood swings" do not affect SWL prediction and confirmed SWL's high temporal consistency. Strong link features, such as the SWL of top friends or significant others, increased prediction accuracy. Our final model incorporated ego features, predicted personality features, and the SWL of strong links. The final model out-performed previous research on the same dataset by 45%.

Keywords: social networking · Facebook · satisfaction with life

1 Introduction

Have you ever Googled "happiness"? If you have, then you have noticed there are about 325 billion results and counting. This is not too surprising as most people consider happiness a desirable goal in life. Happiness has many interpretations; in this paper, we focus on satisfaction with life (SWL). SWL is a component of subjective well-being (SWB), defined by a cognitive judgmental process on how individuals evaluate their lives according to their personal criterion [8].

⋆ corresponding author. Please send inquiries to nmarkuzon@draper.com

Since the 19th century SWL has been studied to identify and improve the quality of life for individuals and nations [23]. From an individualistic prospective, SWL has been used to gain a more robust understanding of mental illness by not only understanding the absence of a pathology but also the presence of happiness [9]. For instance, studies have shown that SWL can predict depression [16], occupational functioning [18], and successful interpersonal relationships [12]. From a community prospective, SWL is used to measure social progress and policy effects. In 2010, David Cameron, the prime minister of the United Kingdom, asked the Office of National Statistics to survey the nation for its life satisfaction as a part of a £2 million per year well-being project [14]. Clearly, the identification and understanding of SWL has risen to national and international attention making it a noteworthy pursuit. The question is, how can we accomplish this identification effectively and effieciently?

With the ubiquity of social media, research in data mining, natural language processing and other computational sciences has dramatically grown [17]. In a 2013 study [11], it was noted that 74% of online adults use Facebook. People are posting about their lives, family, and social interactions making sites like Twitter, Facebook, LinkedIn, etc. gold-mines for data. In other word, these users have already accomplished the tedious and resource consuming work of cataloging their interaction for us. The challenge is how to effectively transform this raw data into knowledge.

In our research, we developed models to predict an individual's SWL from Facebook features and identified indicators and their contributions toward predicting SWL. We took a novel approach by layering machine learning models and incorporating different types of features. We demonstrated the strong correlation between Big 5 personality features and SWL. Big 5 refers to the five broad dimensions of human personality, that include: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism [5]. We predicted Big 5 features and used the predictions in the SWL models. We then combined the predicted personality features with highly correlated static ego features, temporal features, and link features to create a robust model for predicting SWL. We showed that the two-step approach more accurately predicted SWL than one-step models, and outperformed previous research on the same data.

2 Related Work

In recent years, many studies have taken advantage of social media data to evaluate SWL and SWB. These studies have primarily considered ego variables (i.e. personal information such as gender, age, work place, etc.) or link relationships (i.e. how other users influence the ego) to predict SWL or SWB [2, 6, 15, 21].

In a Twitter study, researchers extracted topics and words from tweets to characterize and create a predictive model for SWL. Classic demographic features such as age, sex, and education combined with linguistic features, created the best model for prediction [21]. In another study, Facebook "likes" were used to predict a wide range of private traits and behaviors such as ethnicity, SWL, etc. [15]. This study's model accurately classified some attributes of a user, but

less accurately predicted the numerical label of SWL (Pearson's Correlation R=0.17). Researches explained the less accurate results as SWL's variability caused by "mood swings". "Mood swings" refer to a change in a user's mood over a short period of time.

In addition to static ego features, link features play a role in predicting SWL [2, 6]. Just recently, Facebook researchers analyzed how emotions are spread in the virtual environment. They observed that as positive posts from "friends" were reduced in news feeds that people posted fewer positive updates [6]. In another study of Tweets, researchers determined if assortative mixing took place in an online social network context. Assortative mixing is the tendency for individuals with similar characteristics to favor one another. The researchers concluded that Twitter is assortative, and relationships with more interconnected links were most influential [2].

In our proposed model, we combined previous SWL prediction information with a new layering technique. We incorporated linguistic features from Facebook updates to boost performance, considered temporal features to account for the "mood swing" of users, and incorporated link features by utilizing the SWL of friends as a feature. Our layered model demonstrated the importance of using personality features as an intermediate feature to reduce noise of high dimensional data.

3 Data Description

We used data collected by the myPersonality.org project [15], which contains psychometric test results and Facebook data used for social science research. The dataset contains 101,069 users with SWL scores. There are three feature types: (1) static ego, (2) temporal ego, and (3) link features described below:

3.1 Features and the Target Variable

Static Ego Features. Static ego features belong to a user but do not have timestamps associated. The following are static ego features included in the models:

- **Big 5:** The Big Five features refer to the five dimensions of human personality: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. The Big 5 features were collected through IPIP proxy for Costa and McCrae's NEO-PI-R questionnaire and are numerical variables in the range of [1.0-5.0]. Previous research shows that Big 5, particularly Neuroticism and Extraversion, strongly correlates to SWL [9]. We validated these findings in our data and utilized Big 5 as latent feature to predict SWL.
- **Age:** Reported age of a user.
- **Network Size:** Number of "friends" a user has in his network.
- **Number of Photo Tags:** Number of tagged photos for a user.
- **Relationship Status:** Categorical value representing a user's relationship status.
- **Likes:** Topical decomposition of users Like Data into 600 topics. Topics were extacted using Latent Dirichlet Allocation (LDA) [1].

– **Linguistic Inquiry Word Count (LIWC) Overall:** Linguistic Inquiry Word Count is a text analysis program that counts words into psychologically meaningful categories [22].

Temporal Ego Features. The Facebook status update feature contains temporal information. The status update is free text posted by a user. We used LIWC per post at different time frames for the temporal ego feature. We extracted features reflecting the mood of a user by calculating the LIWC of each update. Daily and weekly averages of LIWC were calculated for each user. We evaluated the potential for "mood swings" [15] by identifying how words used on a daily or weekly basis changed prediction accuracy of SWL. We defined "mood swings" as a change in word usage over time. We hypothesized that features collected closer to the SWL test would be more predictive than features collected further from the test.

Link Features. The third type of features is links associated with users, including friends and couples.

– **Friends:** We utilized the dyads table to calculate mutual friendships of users. We hypothesized that people who share a greater amount of friends were more likely to influence each other. The top 3 friends for each user were identified. Each friend's SWL score was used as a feature to determine whether a friend's "happiness" affects a user. Because not all friends had a true SWL score, we used predicted SWL from Big 5.
– **Couples:** The couples' table was utilized similarly to the friendship table; however, no calculation for rank was required. The SWL score of a significant other was used as a feature to identify how a significant others' "happiness may influence the user. Because not all significant others had a true SWL score, we used predicted SWL from Big 5.

Target Variable: Satisfaction With Life (SWL). The SWL score was the target variable for this study. It was collected from Facebook with the Satisfaction with Life Scale – a 5 item long questionnaire designed to measure global cognitive judgments of satisfaction with one's life [8]. The SWL score is a numerical label ranging from [1.0-7.0] where 1.0 corresponds to highly unsatisfied individuals and 7.0 corresponds to highly satisfied individuals.

3.2 Sample Size

Models had variable sample sizes due to missing values for features. For example, of users with an SWL score, only 85% had Big 5 features and only 4% had LIWC. We calculated the sample size for any particular model by taking the intersection of users who contained all model's features. The sparsity in the features caused models to have drastically different sample sizes. To combat some of these small sample sizes, we chose Facebook features with $n \geq 20,000$ and $R \geq |.50|$. The Static-Ego models ranged from $n = 86,073$ when using Big 5 features to $n = 3,251$ when using LIWC features. Combined Static-Ego models ranged from 11360 to 1160 samples. Combined Link and Static-Ego features had $n = 695$ when friends' SWL were used and $n = 171$ when significant other's SWL was used.

4 Methods

We created data driven supervised learning methods to predict SWL from a set of features extracted from Facebook. In contrast to other models[2, 6, 15, 21], our models considered static ego features, temporal features, link features, and a combination approach for prediction. To reduce noise when combining high dimensional features, we employed a two-step approach of predicting Big 5 as an intermediate feature. To increase performance, we utilized an iterative approach of model building. Starting with the most correlated features from the ego, we expanded our selection to other useful variables, such as link features.

4.1 Model Selection

Although past research [15, 21] predicting SWL used linear regression as a supervised learning model, we utilized Random Forest Regression (RFR) [3]. RFR was used for its interpretability (features can be ranked by importance), non-linear assumptions, efficiency, and accuracy. In this experiment, other methods such as linear regression and support vector regression were explored; however, they did not provide better prediction accuracy and afforded less interpretability than RFR. We employed the scikit-learn implementation of RFR [20]. Mean square error was used as the splitting criterion [19].

Static Ego Models. We used static ego features from each dataset to train Random Forest Regression models to predict SWL. In a two-step approach, we developed models to predict Big 5 from multiple features. Predicted Big 5 scores were then incorporated as features in static ego models. The following summarizes the features for static ego models:

- **Big5:** Big 5 scores collected from a questionnaire [5].
- **FBAttrib:** Age, Network Size, Relationship Status and Number of Photo Tags
- **Likes:** "Likes" of a user as represented by 600-dimensional vector
- **LIWC:** Overall LIWC for a user represented by a 64-dimensional vector
- **Big5.FBAttrib:** Big 5 scores predicted by FBattrib features
- **Big5.Likes:** Big 5 scores predicted by "Likes" of a user
- **Big5.LIWC:** Big 5 scores predicted from LIWC features

Combined Static Ego Models. We combined the best predictors from static ego features into one model to boost performance. When multiple Big 5 predictions were incoporated as features, we used the average for each Big 5 component as a feature. The following summarizes the features for combined static ego models:

- **Combo.Static.1:** FBAttrib features and the mean of Big5.FBAttrib, Big5.Likes, and Big5.LIWC features.
- **Combo.Static.2:** FBAttrib features and the mean of Big5.Likes and Big5.LIWC features.

Temporal Models. Temporal Models tested whether words expressed in Facebook statuses closer to the time of the SWL test had greater prediction accuracy than previous posts. We considered two granularities: daily and weekly statuses. The following summarizes the features for temporal models:

– **Temporal.1:** LIWC derived from Facebook statuses "n" days before SWL test, where n = [1-7]
– **Temporal.2:** LIWC derived from Facebook statues "n" weeks before SWL test, where n = [1-7]

Combined Static Ego and Link Models. Our final models merged Combo.Static.1 with two link features: top 3 friends' SWL and significant other's SWL. SWL scores of link features were predicted from the Big5 model. The following summarizes the features for combined static ego and link models:

– **FBAttrib.Big5.FriendSWL:** Combo.Static.1 features combined with top 3 friends' predicted SWL
– **FBAttrib.Big5.NoFriend:** Combo.Static.1 features of users with top 3 friends: We used this model as a Baseline to determine the lift of the top 3 friends' SWL.
– **FriendSWL:** Top 3 friends' predicted SWL: We use this model to determine the accuracy of the these features by themselves.
– **FBAttrib.Big5.OtherSWL:** Combo.Static.1 Model combined with the significant other's predicted SWL
– **FBAttrib.Big5.NoOther:** Combo.Static.1 features of users with a significant other: We used this model as a Baseline to determine the lift of a significant other's SWL.
– **OtherSWL:** Significant other's predicted SWL: We use this model to determine the accuracy of these features by themselves.

5 Experiment

5.1 Experimental Setting

To evaluate our models we used mean absolute error (MAE) measure [19]. MAE is defined as the average of the absolute errors over n samples, $e_i = |f_i - y_i|$, where f_i is the predicted value and y_i is the actual value.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \tag{1}$$

We evaluated our model by calculating MAE for SWL prediction and comparing it to the MAE of a random model generated from the probability distribution of a sample. The probability distribution function was estimated by interpolating over a 10-bin histogram of the labeled data. Because there are no other experiments that predict SWL from all chosen features, we found the random baseline as a naive but appropriate baseline. To make our models consistent with the qualitative interpretation of SWL scores [10], we consider models with average error rates ≤ 1.0 to be good models.

We also compared our model to a previously discussed model which used linear regression and user likes to predict SWL [15]. We replicated their methods and found that MAE=1.22 ± 0.04 (n=3,920) using the same data in the Likes model. All experimental results were based on 10-fold cross validation.

Table 1. Pearson's R Between Feature and SWL. Some features are averaged over SWL as annotated with *.

Feature	R	# of Samples
agreeableness *	0.988	86073
concientiousness *	0.986	86073
extraversion *	0.997	86073
neuroticism *	-0.998	86073
openness *	0.901	86073
age *	0.249	42264
network size *	0.846	60863
num of tags *	0.596	23197
anger (LIWC)	-0.105	3505
body (LIWC)	-0.106	3505
negative emotion (LIWC)	-0.160	3505
swear (LIWC)	-0.148	3505

5.2 Experimental Results

Table 1 summarizes correlations between features and SWL, confirming the strong correlation between Big 5 and SWL. We observed some correlations between LIWC categories and SWL, signifying linguistic features' utility for predicting SWL. When ego features were averaged over SWL scores, we observed other highly correlated features (age, network size, relationship status, and number of tags), which were subsequently used in prediction models.

Table 2. Static Ego Models

Model	# of Samples	MAE	Random MAE
Big5	86073	0.97	1.58
FBAttrib	9461	1.19	1.34
Likes	3920	1.15	1.57
LIWC	3251	1.16	1.60
Big5.FBAttrib	9242	1.10	1.61
Big5.Likes	3693	1.13	1.56
Big5.LIWC	3251	1.16	1.60

Table 3. Combined Static Ego Models

Model	# of Samples	MAE	Random MAE
Combo.Static.1	1360	1.04	1.64
Combo.Static.2	1190	1.07	1.61

Static Ego Models. After initial feature selection, we created static ego models predicting SWL. Table 2 shows all models perform better than the Random Baseline. All models out-performed a more sophisticated models using "Likes"

features (MAE=1.22) from a previous study [15]. The Big 5 model is the best predictor of SWL (MAE=0.97), and we utilized this insight to create layered models using Big 5 as an intermediate prediction variable. Table 3 shows combination models of static ego features which included predicted Big 5 values as features. Combining ego static features yielded greater accuracy than employing ego features alone (MAE=1.04); however, the Big5 model still out-performed both combination models. This underscores the importance of Big 5 when predicting SWL.

Combined Static Ego and Link Models. Table 4 summarizes the findings when link information was added to the input feature vector. When incorporating the Top 3 Friends features, we observed a slight performance boost (MAE=0.822) in the FBAttrib.Big5.FriendSWL model over the model that did not use link features (MAE=0.827); however, we found that models incorporating link features were significantly better than our previous best model (Big5 with MAE=0.97). This may indicate that not only an individual's situation influences his "happiness" but also the "happiness" of those close to him. Perhaps another explanation for this phenomenom could be that SWL is assortative, where "happy" people gravitate toward "happy" people. When incorporating a significant other's SWL, we observed an even greater performance boost (MAE=0.670). Similar to top 3 friends, the model utilizing significant other's SWL is only slightly better than the model that did not include link features (MAE=0.681). We noted that a significant other's SWL predicted a user's SWL more accurately than his Top 3 friends. This finding is plausible since a significant other is more likely to share in daily life events and may be more influential than "friends" on Facebook.

Table 4. Combined Models of Static Ego and Link Features (above: top-3 friends; and below: significant other).

Model	# of Samples	MAE	Random MAE
FBAttrib.Big5.FriendSWL	695	0.822	1.54
FBAttrib.Big5.NoFriend	695	0.827	1.54
FriendSWL	695	0.865	1.54
FBAttrib.Big5.OtherSWL	171	0.670	1.48
FBAttrib.Big5.NoOther	171	0.681	1.48
OtherSWL	171	0.804	1.48

Temporal Model. Although the temporal models proved predictive of SWL, there was little variance over time. This suggests "mood swings" (expressed through LIWC), do not affect SWL. In particular, Temporal.1's performance showed no significant difference in prediction accuracy when using recent posts (1 Day: MAE = 1.18) versus earlier days (2 Days: 1.18, 3 Days: 1.18, 4 Days: 1.19, 5 Days: 1.19, 6 Days: 1.18, 7 Days: 1.18). Similarly, Temporal.2 showed no significant difference on a weekly scale (1 week: 1.18, 2 weeks: 1.17, 3 weeks: 1.18, 4 weeks: 1.17, 5 weeks: 1.17, 6 weeks: 1.17, 7 weeks: 1.17).

6 Conclusion and Future Work

In this study we created several models to predict SWL. Our findings showed that ego features such as network size, number of photo tags, age, relationship status, likes, and overall word usage (LIWC) can be combined to make a good predictor of SWL. We noted that Big 5 consistently predicted SWL and that reducing variables from high dimensions, e.g. 600-dimensional "Likes", to highly predictive variables, e.g. Big 5, increased the performance of our model.

When using link features, we found a performance boost over the Big5 model. However, when compared to combined static ego feature models, the boost was minimal. This may be attributed to the noise of using a predicted SWL score for friends and couples. If we had the true SWL values for link relationships these models may have shown more lift.

Although LIWC is a good predictor of SWL, the temporal feature of Facebook statuses showed no improvement to our models. This may be attributed to SWL's high internal and temporal consistency as noted in previous research [8]. Because SWL measures a cognitive-judgmental process, it is plausible that "mood swings" expressed by LIWC would not be a large indicator of a user's overall SWL. Another explanation could be that the timeframes were not granular enough to capture the transient mood of a user prior to the SWL test.

Overall, when compared to the Random Baseline, all of our models out performed random prediction by at least 11%. When compared to a linear regression model that used "Likes" features [15], we found our best model was 45% more accurate. We believe that the selection of Random Forest Regression, a combination of static ego features, and "important" link features provided an increase in prediction accuracy.

Our study demonstrated how social media sites such as Facebook contains a set of features useful for predicting private traits. The ability to predict user attributes like SWL may benefit social sciences at the individual and community level. From an individual stand-point, we can create early warnings schemes to identify users who are in distress. For example, SWL could be used as an indicator for identifying issues like depression in students [13] or PTSD in veterans [4]. From a community stand-point, collection of SWL from social media can provide an efficient evaluation for public wellness. For government entities like the European Union, this could mean saved efforts and costs for collecting and processing international surveys on SWB[14].

A major limitation of this study was sparse features. Some features (e.g. number of groups) correlated highly with SWL ($R = -0.678$) but were not well populated, and therefore were not utilized as a predictive feature. We suggest future work to focus on fine-tuning feature collection and selection. We also suggest using link relationships as predictors of SWL. In our models, we saw promising results from link features (MAE=0.67), however the sample size (n=171) was relatively small.

7 Acknowledgments

This research was partially supported by the Draper Laboratory internal Research and Development funding.

References

1. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022.

2. Bollen, Johan, et al. “Happiness is assortative in online social networks.” Artificial life 17.3 (2011): 237-251.
3. Breiman, Leo. “Random forests.” Machine learning 45.1 (2001): 5-32.
4. Bryant, Richard A., et al. ”Posttraumatic stress disorder and psychosocial functioning after severe traumatic brain injury.” The Journal of nervous and mental disease 189.2 (2001): 109-113.
5. Costa Jr, P. T., and Robert R. McCrae. “Neo personality inventoryrevised (neo-pi-r) and neo five-factor inventory (neo-ffi) professional manual.” Odessa, FL: Psychological Assessment Resources (1992).
6. Coviello, Lorenzo, et al. “Detecting Emotional Contagion in Massive Social Networks.” PloS one 9.3 (2014): e90315.
7. Diener, Ed. “Subjective well-being: The science of happiness and a proposal for a national index.” American psychologist 55.1 (2000): 34.
8. Diener, E. D., et al. ”The satisfaction with life scale.” Journal of personality assessment 49.1 (1985): 71-75.
9. Diener, Ed, Shigehiro Oishi, and Richard E. Lucas. “Personality, culture, and subjective well-being: Emotional and cognitive evaluations of life.” Annual review of psychology 54.1 (2003): 403-425.
10. Diener, Edward. ”Understanding scores on the satisfaction with life scale.” Retrieved August 8 (2014): 2009.
11. Duggan, Maeve, and Aaron Smith. ”Social Media Update 2013.” Pew Research Centers Internet American Life Project RSS. Pew Research Center, 30 Dec. 2013. Web. 07 Aug. 2014.
12. Furr, R. Michael, and David C. Funder. ”A multimodal analysis of personal negativity.” Journal of personality and social psychology 74.6 (1998): 1580.
13. Frisch, Michael B., et al. ”Predictive and treatment validity of life satisfaction and the quality of life inventory.” Assessment 12.1 (2005): 66-78.
14. GOV.UK. ”Wellbeing: Introduction to Subjective Wellbeing Datasets.” Research and Analysis. Cabinet Office, 27 Mar. 2013. Web. 08 Aug. 2014.
15. Kosinski, Michal, David Stillwell, and Thore Graepel. ”Private traits and attributes are predictable from digital records of human behavior.” Proceedings of the National Academy of Sciences 110.15 (2013): 5802-5805.
16. Lewinsohn, Peter M., J. Redner, and J. Seeley. ”The relationship between life satisfaction and psychosocial variables: New perspectives.” Subjective well-being: An interdisciplinary perspective (1991): 141-169.
17. Manyika, James, et al. “Big data: The next frontier for innovation, competition, and productivity.” (2011).
18. Marks, Gary N., and Nicole Fleming. ”Influences and consequences of well-being among Australian young people: 19801995.” Social Indicators Research 46.3 (1999): 301-323.
19. Rice, John. Mathematical statistics and data analysis. Cengage Learning, 2006.
20. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
21. Schwartz, Hansen Andrew, et al. “Characterizing Geographic Variation in Well-Being Using Tweets.” ICWSM. 2013.
22. Tausczik, Yla R., and James W. Pennebaker. ”The psychological meaning of words: LIWC and computerized text analysis methods.” Journal of language and social psychology 29.1 (2010): 24-54.
23. Veenhoven, Ruut. The study of life-satisfaction. 1996.