Predicting a Happier Place

Lisa Wang Karen Wang Grant Means
{lisa1010,kwang37,ghmeans} @ stanford.edu

I. Introduction

In communities with higher well-being, people live longer and more productive lives, and local economies flourish. In an effort to improve understanding of what leads to vibrant communities and fulfilling lifestyles, our project explored community-level factors and their significance for determining well-being in Americas largest metropolitan areas. We hypothesized that the frequency of points of interests (POIs) in a community as well as basic macro-level factors were sufficiently accurate indicators of well-being. Thus, the input to our algorithm was POI frequencies and economic/social factors for each metro area. We then used Logistic Regression, Extra-Trees Classifier, and Support Vector Machines to output a prediction of whether a metro area is of high or low well-being.

II. Prior Work

In 2013, Schwartz and Eichstaedt et. al. from the University of Pennsylvania and Michigan State University predicted the well-being of 1,300 US counties using sentiment analysis on Twitter data. They used the heuristic of life satisfaction (LS) as measured by random phone surveys in response to questions such as how satisfied are you with your life?, with choices ranging from very dissatisfied to very satisfied. We think that it would be interesting to build upon the implications raised by this paper that location is a fundamental underlying factor in peoples overall happiness or satisfaction with life [3].

In 2012, Quercia and Ellis et. al. studied the relationship between sentiment expressed in tweets and community socio-economic well-being for Twitter users throughout London. They found that tweet sentiment for individuals in a community and the socio-economic well-being of the community were highly correlated. The usefulness of social media features in determining well-being could help inform future feature construction for our wellbeing predictions [4].

In 2005, Joachims explored the use of SVMs for learning text classifiers from examples. He found that SVMs achieve substantial improvements over other methods of classification and behave robustly over a variety of different learning tasks. This indicates that SVMs are a state-of-the-art method and validates our decision to use SVMs for classification. Given these results, it makes sense that an SVM was our most accurate method for classification [5].

In summary, using Twitter data to gain insights into communities was a clever approach to determining well-being. Based on the available literature regarding well-being, it appears that the majority of prior work was done by hand using traditional survey methods or aggregating previous research. The only existing well-being applications we could find for machine learning involved Twitter data, so our features appear to add a unique spin on the issue of community factors in well-being.

III. Datasets

OpenStreetMap (OSM) is a collaborative, open-source alternative to Google Maps [1]. The data tags points of interest (POIs), marked by their longitude and latitude coordinates, with their names and functions.

The Gallup-Healthways Index is a collaborative project that provides data on the community, financial, physical, purpose, and social well-being scores, as well as overall city satisfaction of the 107 largest metropolitan communities across America [2]. Each community is assigned an index, or score, for each well-being category based on phone interviews with random samples of people from each community. Gallup defines scoring high in each category to mean:

- **Community:** Liking where you live, feeling safe and having pride in your community.
- **Financial:** Managing your economic life to reduce stress and increase security.
- **Physical:** Having good health and enough energy to get things done daily
- **Purpose:** Liking what you do each day and being motivated to achieve your goals.
- **Social:** Having supportive relationships and love in your life.

Well-being scores took on a floating point value between 5.0 and 8.0. Overall city satisfaction took on a percentage value indicating what proportion of respondents interviewed felt generally satisfied with the city area they live in. For our project, we defined high well-being as having a high score in one or more of these indexes. Although Schwartz and Eichstaedt et. al. used the LS heuristic for well-being as discussed above, this would have involved a non-trivial challenge on our part of acquiring the coordinates for more than one thousand US counties to extract OSM data. We decided that the Gallup data, which covers a smaller but more well-defined set of large metropolitan areas, was more suitable for our purposes.

Community statistics were culled from various governmental, commercial and NGO sources, including the Tax Foundation [6], [7], U.S. Department of Commerce [10], U.S. Census Bureau [9], and *The New York Times* [8].

IV. Problem Approach & Features

Initially, we sought to predict the community well-being score (as defined in the Gallup data) for each area by training a set of regression models. The features included frequencies of POIs which we believed to be correlated with well-being. We extracted the 97 most frequent tags from OSM, which included leisure and amenity resources such as fitness centers, beach resorts, pubs, arcades, hot springs, taxi stands, etc. In order to tease out meaning from values between metro areas with populations that span well over an order of magnitude, we included frequency per capita by normalizing POI counts based on the communitys population size: for each community, we computed a scaling factor of its population size divided by the minimum population size over all communities, and multiplied each POI frequency by said scaling factor.

However, after evaluating our initial results, we decided to pivot from regression models to classification models instead. Moreover, the Gallup data had a relatively small range (ex. social well-being scores only ranged from 5.8 to 6.4, discretized into intervals of 0.1), which made it difficult to perform regression on. We also expanded our feature set to attempt to capture more macro-level phenomena of a community. OSM classifies POIs into larger categories of sustenance, education, transportation, financial, health care, culture, and sport. We postulated that the counts of POI categories may be just as important as the counts of individual POIs in predicting a communitys well-being, and thus included these in our feature set. We also included additional features representing macro-qualities of cities such as regional classifications, coastal proximity, and geography for a total of 115 features per community. A sample of our data is given in Table I.

TABLE I
Excerpt from dataset after feature normalization

Community	Social Well-Being score	`taxi stand`	`common area`	`city tax rate`
Reno, NV	6.4	1.11	47.86	0.00
Youngstown-Warren-Boardman, OH-PA	5.8	0.00	4.16	2.75

V. Data Collection & Preprocessing

The Gallup data provided us well-being heuristics for the 107 most populous metropolitan communities in the US, where communities are defined as groups of one or more cities. While extracting data from OSM, we ran into errors with the extraction API, and were only able to extract POI counts for 104 different communities for our regression models.

For our classification approach, we decided to split our targets into the top $\frac{1}{3}$, middle $\frac{1}{3}$, and bottom $\frac{1}{3}$ based on each individual well-being score. We then discarded the middle, and used the top and bottom for a total of 68 training examples to perform binary classification. Since data points in the middle are very similar, his method provides a clearer distinction between data points of high and low well-being compared to simply splitting the data in half.

VI. Regression Models

A. Linear Regression

We started with linear regression since it is the most straight-forward regression model. It minimizes the squared error by performing the Ordinary Least Squares algorithm. The minimization problem can be described as:

$$min_\theta ||X\theta - y||_2^2$$

where X is the mxn design matrix containing the features of the m training samples, θ is the $nx1$ vector of weights and y is the $mx1$ vector of targets.

B. Epsilon Support Vector Regression

SVR uses support vector machines (SVM's) to perform regression by minimizing the norm of the weights subject to approximating the pairs (x_i, y_i) with ϵ precision. The advantage of using SVRs included the possibility of using kernels; we used linear and RBF.

C. Other Models

We also tried ridge regression, lasso regression, polynomial regression, and epsilon support vector regression with polynomial kernel, but found these to perform worse than the models above.

VII. Regression Evaluation Methods

A. Train-Validation-Test Split

For the regression approach, we set aside 90% of our data to split into a 80/20 train and validation set, which we used to improve our models. The remaining 10% was reserved as test data which we did not touch until we tuned our models based on our train and validation sets.

B. Regression Metrics

Since we were working within a regression model, we used mean-squared error (MSE) as well as the explained variance score (EVS) to compare the performance of our models on the training as well as the validation set.

1) Mean-Squared Error: Let m be the number of samples, let y_i be the true target of the i-th sample, and let $\hat{y}_i$ be the predicted value of the i-th sample. Then, the mean-squared error is defined as:

$$MSE(y, \hat{y}) = \frac{1}{m}\sum_{1}^{m}(y_i - \hat{y}_i)^2$$

2) Explained Variance Score: This score measures the proportion of variance created by the regression model to the true variance of the data set. A score of 1.0 is perfect and lower scores are worse.

$$EVS(y, \hat{y}) = 1 - \frac{Var(y - \hat{y})}{Var(y)}$$

VIII. Regression Feature Selection

We used three different methods to perform feature selection for our regression models. We hoped that feature selection would help us determine the most significant features to improve performance and find the amenities in a community most correlated with well-being.

A. Single Feature Regression

We ran all our regression models on the training data separately for each feature to determine the most important independent features. However, this approach does not capture e.g. the covariance between multiple features.

B. Recursive Feature Selection (RFE)

This method selects features by starting with the full feature set and recursively reducing the feature set until the desired number of features is reached. It repeatedly trains the given estimator on the data, eliminating the features with the lowest weights. Hence, it is very similar to backward search.

IX. Regression Results

After performing feature selection using single feature regression, we reduced the number of features to 12. The 12 features include among others the tags "doctors", "fire station" and "park". We computed the MSE and EVS for both training and validation sets. Our results for regression are shown in Table II.

TABLE II
Evaluation of Regression Models

Model	MSE train	EVS train	MSE valid	EVS valid
Linear regression	13.50	16.83	0.21	0.34
SV regr. linear	15.86	19.98	0.07	0.22
SV regr. RBF	12.94	17.63	0.24	0.31

X. Classification Models

A. Logistic Regression

Logistic regression attempts to find parameters θ that best fit the model:

$$y = 1 \text{ if } \theta^T x > 0$$
$$y = 0 \text{ otherwise}$$

The parameters are found using maximum likelihood estimation with the probability distribution given by the logistic function:

$$g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

B. Support Vector Machines

SVMs attempt to find a separating hyperplane between two classes of data by fitting the model:

$$y = 1 \text{ if } w^T x + b > 0$$
$$y = 0 \text{ otherwise}$$

It attempts to give the biggest buffer against error by optimizing the primal objective defined as:

$$min_{\gamma,w,b} \frac{1}{2} \|w\|^2$$
$$s.t.\ y^{(i)}(w^T x^{(i)} + b) \geq 1, i = 1, ..., m$$

However, in order to solve the problem efficiently, SVMs convert the primal objective into its dual form and leverages kernels which map lower-dimensional data into higher-dimensional spaces in which data can be more easily separated. For our study we used linear and (Gaussian) radial basis function, or RBF, kernels.

C. AdaBoost

AdaBoost is short for adaptive boosting, and is an ensemble method that uses many other types of learning algorithms (weak learners) and computes a weighted sum representing the final output of the boosted classifier. A boost classifier is of the form:

$$F_T(x) = \sum_{t=1}^{T} f_t(x)$$

where each f_t is a weak learner that takes x and returns a real value indicating its class. As long as the performance of each weak learner is better than random guessing, the final classifier can be shown to be a strong learner.

D. ExtraTrees

Decision trees are used in supervised classification problems to create a model that predicts target values using decision rules. ExtraTrees is short for extremely randomized trees, and is an ensemble method that relies on randomly splitting decision tree nodes. The ExtraTrees algorithm is more computationally efficient than Random Forests and produces a smoother decision boundary due to the randomization.

XI. Classification Evaluation Methods

Since we only had 68 training examples, we decided to use k-fold cross validation rather than make a standard train/validation/test split. We chose $k = 20$ in order to leave out less data each time. Moreover, since our evaluation heuristics of F1 scores, precision, recall, and accuracy tend to have large variances, a higher k produces more trials to average over and thus higher confidence in our evaluation results.

We standardized the data through mean removal and variance scaling such that each feature has mean 0 and variance 1. The SVM methods expect the mean for each feature to be 0 and the same variance for all features makes sure that each feature has equal importance. We performed five classification algorithms on our data set, using three different kernels for the support vector machines. Running the algorithms without feature selection produced results that were not much better than random classification. However, that was not a surprise since we had 115 features. Hence, we performed feature selection by scoring each feature with the statistical ANOVA

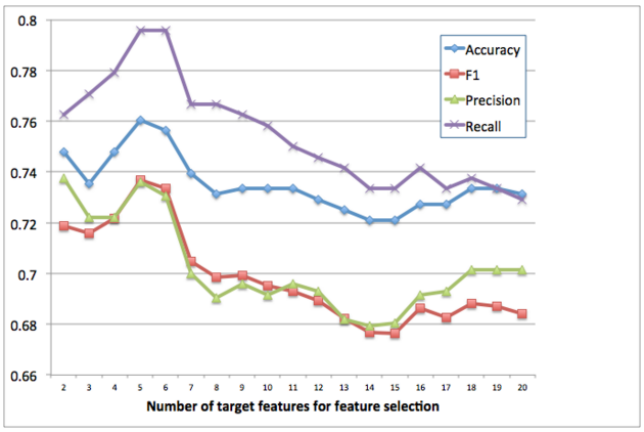

Fig. 1. Evaluation of number of features

F-Score and chose the best k features. We tried out different k values in the range between 2 and 20 and found 5 to be the best. After feature selection, we ran all classification algorithms on each data set of the six data sets.

To evaluate our approaches, we computed the accuracy, precision, recall and the resulting F1 Scores for each experiment. Accuracy in the binary classification context is just the percentage of predictions that were correct, so it is computed by

$$\frac{TP + FN}{TP + FP + TN + FN}$$

XII. Classification Feature Selection

A. Choosing the Number of Features for Feature Selection

To find the best number of features for feature selection, we did a search using SVM with RBF kernel over feature numbers in the range between 2 and 20. The best accuracy and F1 score was achieved when we selected the 5 best features according to their ANOVA scores (Fig. 1). Decreasing or increasing the number of target features impacted the accuracy and F1 scores significantly.

B. Feature Selection Rankings

We performed feature selection on each dataset separately. Table III shows the five best features for each dataset. Interestingly, the most indicative features were exactly the same for the community well-being and the overall city satisfaction datasets, which is a sign that community well-being might be the most important indicator of overall satisfaction. In addition, the selected features for these two datasets consisted mostly of non-physical features such as college graduation rates or GDP growth rather than amenities. In contrast, social well-being is mostly correlated with amenities and physical features such as taxi stands, video arcades, clubs and common public areas. Even though we cannot make any conclusions about causation, this information shows that the most indicative features are very different between the datasets. While physical features are good indicators for physical and social well-being, they are less significant indicators for community well-being and overall city satisfaction. Hence, just by looking at the materialistic features of a city, we might not get a good idea of how happy the people living there actually are.

TABLE III
Feature Selection Rankings for each index

Index	Financial	Physical	Purpose
Rank 1	bar	community center	pharmacy
Rank 2	playground	waste basket	water park
Rank 3	college grads	dentist	stadium
Rank 4	coastal	fitness center	city tax
Rank 5	gdp per capita	pop. growth	pop. growth
Index	Social	Community	Overall City Satisfaction
Rank 1	taxi stand	college grads	college grads
Rank 2	video arcade	city tax rate	city tax rate
Rank 3	club	gdp per capita	gdp per capita
Rank 4	common area	gdp growth	gdp growth
Rank 5	city tax	pop. growth	pop. growth

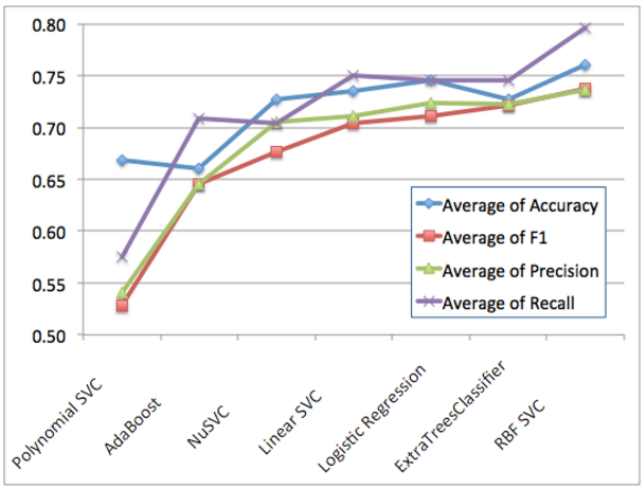

Fig. 2. Evaluation of learning algorithms

XIII. Final Results

A. Comparing Algorithms

We started with logistic regression as our baseline and tried support vector classification with three different kernels as well as ensemble methods, including AdaBoost and ExtraTrees. logistic regression already performed well with an average F1 score and accuracy above 0.7 (*Fig. 2*). Only two other models we tried were able to improve the performance of logistic regression, with support vector classification using an RBF kernel at the top, reaching an accuracy of 0.76 and an F1 score of 0.74. Since SVC with an RBF kernel worked better than SVC with a linear kernel, this could be a sign that there is more complexity in our data and higher-dimensional functions are needed to model it. The ensemble classifier ExtraTrees had higher F1 than logistic regression, but lower accuracy. Since ExtraTrees is a more complex classifier and did not outperform logistic regression, it is usually better to choose the simpler one.

B. Comparing Datasets

We also computed the scores of SVC with RBF for each dataset (*Fig. 3*). As the chart shows, the scores varied signifi-

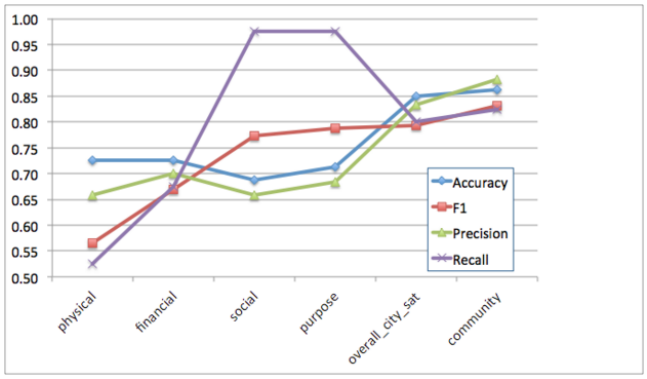

Fig. 3. Evaluation of datasets

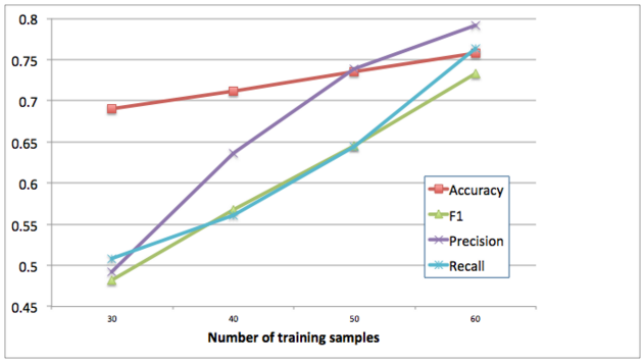

Fig. 4. Evaluation of training set size

cantly between the datasets. The best results were achieved for the Community dataset, with the highest F1 score and balanced precision and recall. For our application, we desire precision and recall to be balanced, especially since both classes are equally important. Interestingly, the datasets Social and Purpose had very high recall, but lower precision compared to the other datasets. This could indicate that the positive class of higher well-being was easier to predict than the the negative class for that particular dataset.

C. Increasing training set size

For evaluating sample set sizes we used k-fold cross validation. Since we began with 30 training examples on which we decided that $k = 10$ was the maximum number of folds that made sense. The trend of our evaluation scores with training set size suggested that our results grew more reliable with a bigger training set (*Fig. 4*). Note especially that recall got closer precision, which is desirable since both are important. For a training set size of 30, the accuracy was much higher than precision or recall. This is because accuracy takes class size into account, and k-fold does not necessarily give the same distribution of happy/unhappy classes.

XIV. Conclusion

Our work showed that predicting well-being of communities in the U.S. using a combination of OpenStreetMap and municipal data as features is a promising approach, which could potentially be further improved by increasing the training set size. By discretizing the problem into a binary classification problem, we simplified the task, but were able to make more concrete progress as measured by accuracy and F1 scores. Feature selection turned out to be very useful for improving our results and also gave us interesting insights about our data. Out of the five well-being categories, community well-being seemed to be most correlated with overall city satisfaction, and they both shared the same set of most significant statistical features. However, POIs were helpful to identify social and physical well-being. We were able to achieve good results for classifying high vs. low well-being scores, with the best results for community well-being. Overall, our work is useful for municipal leaders to get insights into the different aspects of well-beings in their communities and can help indicate well-being levels when polling data does not exist. OpenStreetMap and municipal data are available for many countries around the world, and the data is being expanded everyday. Hence, by combining these globally available datasets, we have experimented with a new approach to predict and analyze well-being using machine learning techniques, that could be further generalized to study well-being across the world.

XV. Future Work

Some of the limiting factors for our exploration of happiness were our intuition for which features are most highly correlated with wellbeing and how much data collection and preprocessing was required to construct a sufficiently robust set of features. A more in-depth exploration of happiness would go beyond these limitations to employ all OSM tags for each metro area and expand statistical features to include a much larger corpus of macro-level community information. During data preprocessing and feature definition we decided to normalize our features linearly based on population size. However, the actual data might not scale linearly, and we would need to do more research on whether this was an accurate assumption. For our classification problem, we could have tried not discarding the middle one-third of training examples. We also could have tried more hyperparameter tuning in our learning algorithm evaluation, especially for SVM such as using grid-search.

In the process of our study we were able to create a novel dataset that combined municipal with OSM data, which we can publish to future researchers can use it. Finally, we could expand our study to include communities outside of the US, since an implication of our study was being able to predict well-being reliably in communities even when there is no polling data. OpenStreetMap data exists over the entire world and is continuously updated, which would allow us to assess the happiness of communities that are not covered by Gallup-Healthways. Since individual surveys are time-consuming and expensive, this could allow us to extend our knowledge of the worlds overall well-being at a large-scale.

References

[1] Main Page. Retrieved November 16, 2015, from wiki.openstreetmap.org/wiki/Main_Page

[2] Gallup-Healthways Well-Being Index. Retrieved November 16, 2015, from www.gallup.com/poll/106756/galluphealthways-wellbeing-index.aspx
[3] Schwartz, Andrew H., Eichstaedt, Johannes C. et. al. (2013). *Characterizing Geographic Variation in Well-Being using Tweets*
[4] Quercia, Daniele ,et al. Tracking Gross Community Happiness From Tweets. The 2012 ACM Conference on Computer Supported Cooperative Work. Association for Computing Machinery. Seattle. 2012. ¡http://dl.acm.org/citation.cfm?id=2145347¿.
[5] Joachims, Thorsten. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Springer. New York. 2005. ¡http://link.springer.com/chapter/10.1007/BFb0026683¿.
[6] Walczak, Jared. "State Individual Income Tax Rates and Brackets." The Tax Foundation. Washington, D.C. 2015. ¡http://taxfoundation.org/article/state-individual-income-tax-rates-and-brackets-2015¿.
[7] Henchman, Joseph, Sapia, Jason. "Local Income Taxes: City- and County-Level Income and Wage Taxes Continue to Wane." The Tax Foundation. Washington, D.C. 2011. ¡http://taxfoundation.org/article/local-income-taxes-city-and-county-level-income-and-wage-taxes-continue-wane¿.
[8] "Cities with the Most College-Education Residents." *The New York Times*. New York. 2012.
[9] "Table 2. Annual Estimates of the Population of Combined Statistical Areas: April 1, 2010 to July 1, 2012". 2012 Population Estimates. United States Census Bureau, Population Division. March 2013. ¡https://www.census.gov/popest/data/metro/totals/2012/tables/CBSA-EST2012-02.csv¿.
[10] "News Release." United States Department of Commerce: Bureau of Economic Analysis. September 2014.
[11] Pedregosa, Fabian, et al. "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research. October 2011.