An Evaluation of Micro-blogging as a Regional Mental Health Indicator

Brandon Olivier, Nicholas Sundin

Department of Computer Science, University of Texas at Austin,
2317 Speedway, Stop D9500 Austin, TX 78712

Vocabulary in Twitter posts was used to score the sentiment for areas in the United States. This score, or happiness index, was aggregated by zip code using the inherent geographical information present in some Twitter posts. Conclusions regarding correlation and usefulness of the happiness index were then gained by comparing it against other available regional metrics. Based around this calculated happiness index, a model was then created that has shown reasonable predictions of suicide rates in California given an aggregate per-capita income and happiness index.

Introduction

The idea that our speech conveys more information than just semantic meaning is an old one. In Psychopathology of Everyday Life, Sigmund Freud describes Fehlleistungen, a "slip of the tongue", or "Freudian slip." (*2*) Freud's research into what we say by mistake has been succeeded by research into how we say what we say. One may suspect that studying the use of emotionally charged words would give the best idea of how people are feeling, but research by James Pennebaker has found that "pronouns in particular have been found to strongly cor-

relate with health improvements." (*1*) The idea that emotionally charged words are the key to understanding the mental health of speakers is wrong. Pennebaker extensively uses a new classification for words: particles. A particle is any functional word that requires other words to derive meaning. For instance, articles, prepositions, and conjunctions are particles. In explaining the importance of particles, Pennebaker says "[t]o use a pronoun requires the speaker and listener to share a common knowledge", people need to relate and have an understanding of the other person; we see this same thing where "informal settings presuppose a shared frame of reference . . . the discerning particle user must have some degree of social and cognitive skill." (*1*)

Much of the speech used in everyday life can be analyzed in terms of particles. Because "[i]n the English language there are fewer than 200 commonly used particles, [and] they account for over half of the words we use." (*1*) Particles are ideal for analyzing the mood or feelings of someone speaking. Since much of the content of our everyday speech is dictated by the situation we are in, merely analyzing the content of a conversation will not provide as clear of an insight into the psyche of the speaker. Particles, on the other hand, are an element of the style of speech a person is using, and as such reflect their attitudes at that time, making them ideal units for sentiment analysis, the use of natural language processing to identify and extract subjective information in source materials.

Twitter, a social media site limiting posts to 140 characters, was created in 2006 and immediately surged in popularity. Twitter boasts more than 350 million messages, known as tweets, published every day as of 2012. (*5*) On a cursory glance one would think that tweets could not be used for much analysis because of their character limitations. However, based on Pennebaker's The Secret Lives of Pronouns and other research materials, many believe tweets "can be combined to build a larger picture of the user posting them." (*1*)

Microsoft research, in 2013, published a study wherein they collected data from Twitter and subjects diagnosed with depression. They used the tweets from the individuals to train a computer to detect signs of depression based solely on the person's Twitter feed. According to that study, some aspects of a user's feed that may indicate an onset of depression include: "decrease in social activity, raised negative affect, . . . , heightened relational and medicinal concerns, and greater expression of religious involvement." (*6*)

Method

Our research builds on the idea that Microsoft Research discusses: we want to analyze tweets from users on Twitter and assign them a grade on their mental health based on their tweets. We can look at tweets from each user, assess each tweet, then assign a score to the user based on the content of their tweets. We then want to aggregate users based on their zip code, yielding an average happiness index for each zip code. After that, we will use other data sources, such as weather, government grants, and suicide rates to try to build a model that will be able to predict our derived ratio independent of tweets.

In 2011, Twitter altered their terms and services and it is now a violation to share full sets of data. Ids are still permitted to be shared, but the actual collection of the full information must be done on a per use basis. For our dataset, we got a list of user ids for twitter users and wrote a python script to gather tweets and save them in a json object that we could later use. For each user, we gathered a collection of tweets, inferred a user's location from the geographic data encoded into tweets, and created a score for each user. The score is based on usage of words. We used 2 modified versions of positive and negative words lists that we amended to contain some internet slang abbreviations and particles. We split each tweet and compared the frequency of words from each list as follows:

```
def calculate_happiness_ratio(tweet_content):
    words = word_counter(tweet_content)
    score = 0
    total_words = 0
    for word, count in words.items():
        if word in positives:
            total_words += count
            score += count
        elif word in negatives:
            total_words += count
    if total_words == 0:
        return None
    else:
        return float(score) / total_words
```

where `word_counter` is a function that creates a python dictionary with each word and the corresponding times that that word is used in the body of the tweet. After assigning a score to each tweet, we reduce to users, and then to zip codes.

Twitter encodes location data as latitude and longitude. As such, we had to discover which zip code any arbitrary latitude and longitude pair falls in. For that, we turn to MongoDB, which has built in support for geolocation. The first step in the geolocation process is to acquire a file detailing the shapes of every zip code from the US government, the exact details of getting the shapefiles is available in the GitHub for the project. (*8*) We used a library called US-Atlas from Mike Bostock toget the shapefiles and We used a tool called ogr2ogr to convert theshapefiles into the JSON format that MongoDB understands. Once the shapes are stored in MongoDB,

queries such as seen below.

```
db.collection.findOne(
  { geometry:
    { $geoIntersects:
      { $geometry :
        { type: "Point",
          coordinates: tweet.coordinates
        }
      }
    }
  }, callbackFunction
);
```

Here, the zip code can be handled arbitrarily in a `callbackFunction`. Each user was run through a query like this to assign them a zip code. After we processed all the data, it was reduced to a CSV file containing a zip code and a corresponding happiness score.

To find data to correlate our scores to, we went to enigma.io, a website devoted to obtaining large datasets and making them publicly available and searched for data indexed by zip code. One of the more interesting datasets we used was average income indexed by zip code. We plotted the data in Tableau and didn't see a correlation between income and happiness index. There are some zip codes that abide by the maxim "money doesn't make you happy", but others that do not. They show that if you make more money, you are more likely to be happy.

In addition to these methods, we used SQL developer and Oracle Data Miner to to process happiness index, suicide rate, and income columns using the decision tree, naive Bayes, and

SVM (support vector machines) to analyze our data. We found no statistically significant link between the three columns using these methods.

Results & Discussion

We created a histogram to see the suicide rate of California zip codes plotted against their corresponding happiness index. We obtained the suicide rate by dividing the number of suicides that occur in a given zipcode, as provided by the California statistics, and we divide that by the total number of deaths in a zipcode. As the histogram shows, there is a correlation between happiness and suicide: as the happiness index decreases, so does the suicide rate for that zip code.

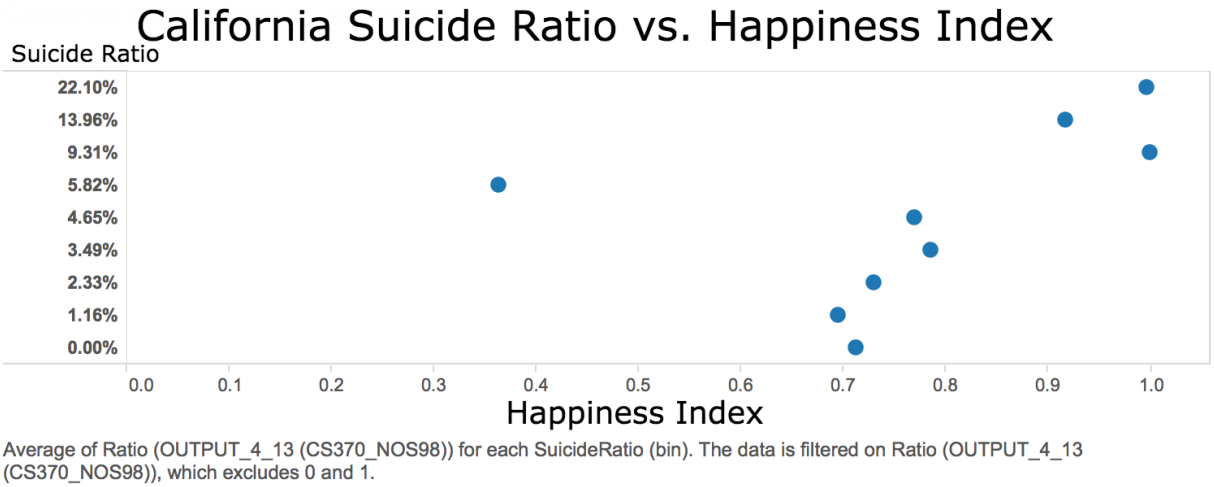

Figure 1: Suicide vs. Happiness

To resolve some discrepancies, we modified what we were looking at. In figure 2, we divide the average income by the happiness index to offset the effects of poverty on mental health. That yields a value which represents the income per happiness unit. We call this a happiness

normalized income (HNI). So if a person lives in place A, with an HNI of 120,000, and is considering moving to a place with an HNI of 100,000, then to maintain their level of happiness, they only need to make $100,000 per year, as opposed to $120,000. Thus HNI is a value that represents how much money one needs to maintain a certain level of happiness. We then plot the suicide rate by the HNI in a new histogram. The suicide rates in figure two are aggregated by average and collected into bins for display purposes. The points in figure 2 are colored to represent the happiness ratio.

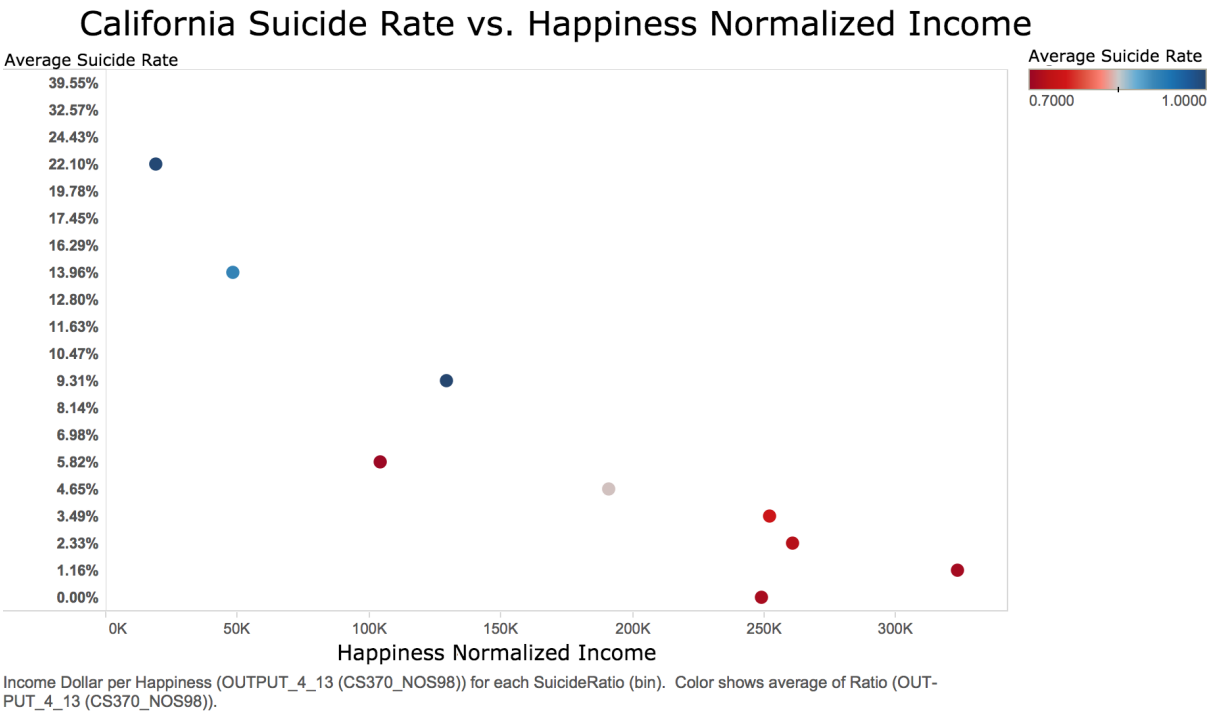

Figure 2: Suicide vs. Happiness Normalized Income (HNI)

There is a correlation between suicide rate and HNI. Figure 2 shows that if one requires more money to be happy in a particular area, then that area has fewer suicides. For instance, an area with a HNI value of 80K should have an higher suicide rate than a zip code that has an HNI of 100K.

We can conclude from these graphs that people in low income areas and areas that have higher measured happiness indices are more likely to commit suicide. While it may seem contradictory that areas with higher measured happiness have a higher incidence of suicide, it would make sense if online posts could be considered a healthy venting of frustration.

We then used this correlation to make predictions about the suicide rate in different areas. Rather than make complicated formulas, we elected to use a simple linear fit line to make predictions. It will be less accurate for many values, but it should still give overall correct results. Many of the predictions are muted in the middle of the spectrum, but on the whole, it seems to give a decent prediction of suicide rates in different zip codes in California.

We also found an unusual trend in a graph comparing happiness to income. For incomes less than $800K, the happiness index is relatively flat; people are just as happy whether they make $80K or $200K. However, at $800K yearly income per capita (per person in the household), something interesting happens: the happiness index plummets, then recovers. One of the more interesting things about the trend is that only one of the bars past $800K is only one zip code (the highest bar is just 10023, a zip code corresponding to an area in the middle of Manhattan, near Central Park and the Upper West Side). Since the others all include multiple zip codes, we believe that they are not outliers or mistakes.

We predict two possible causes for that. One potential cause is that once a person reaches $800K in yearly salary, they enter a new social circle that values money much higher than their previous social circle. That makes them less happy, if they are dependent on their money for happiness. They are now at the bottom of their social circle in terms of monetary wealth.

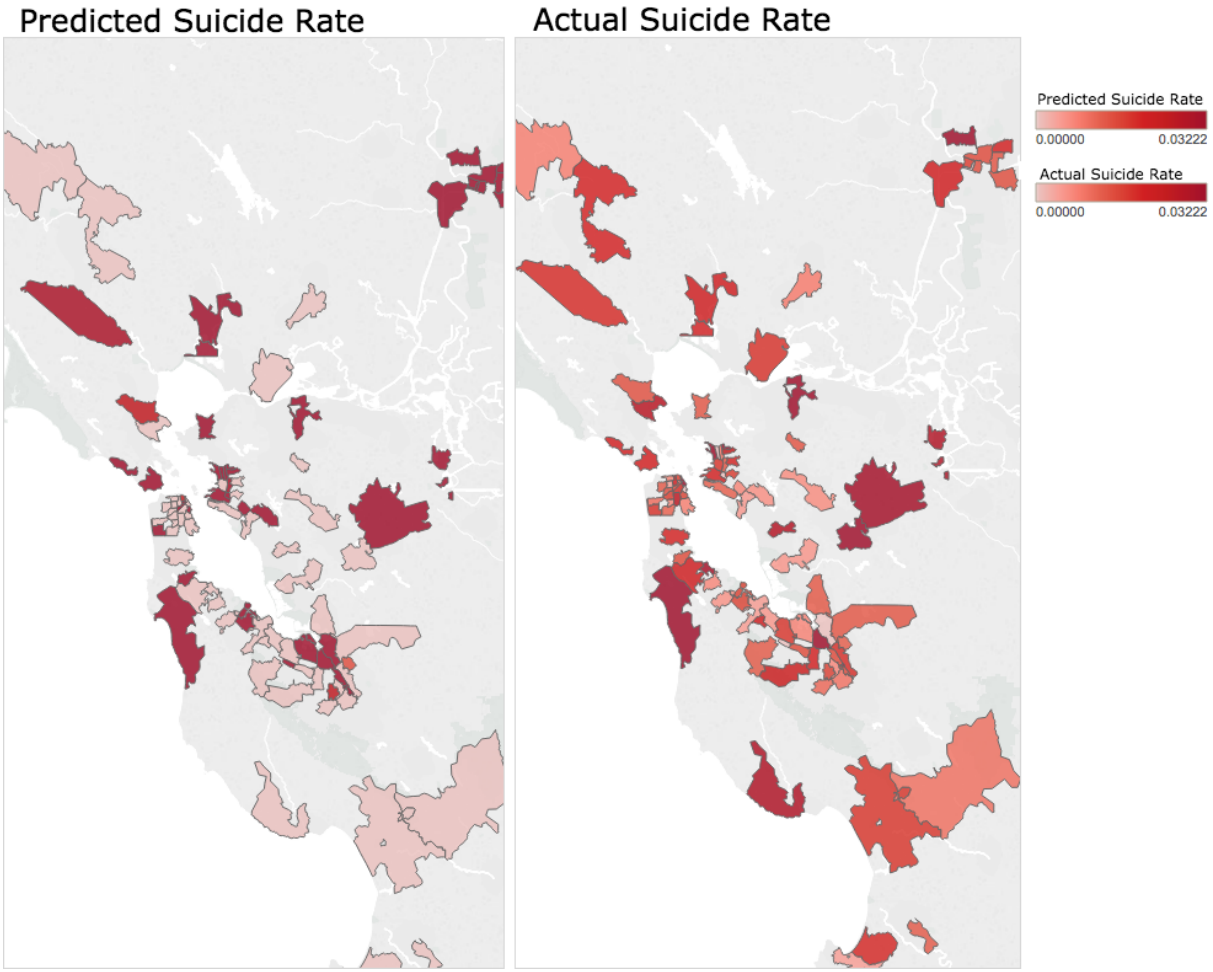

Figure 3: Predicted vs. Actual Suicide Rate

Another potential cause is that there is some single area that is an outlier and is requiring a higher income for happiness. We believe this to be an unlikely cause because it would be very odd that there is some area where one is affected by only earning $800K. That is a lot of money. To test that hypothesis, we filtered the happiness ratio to exclude some areas in New York, as that's the most expensive place to live in the United States. When excluding New York, one gets the following graph.

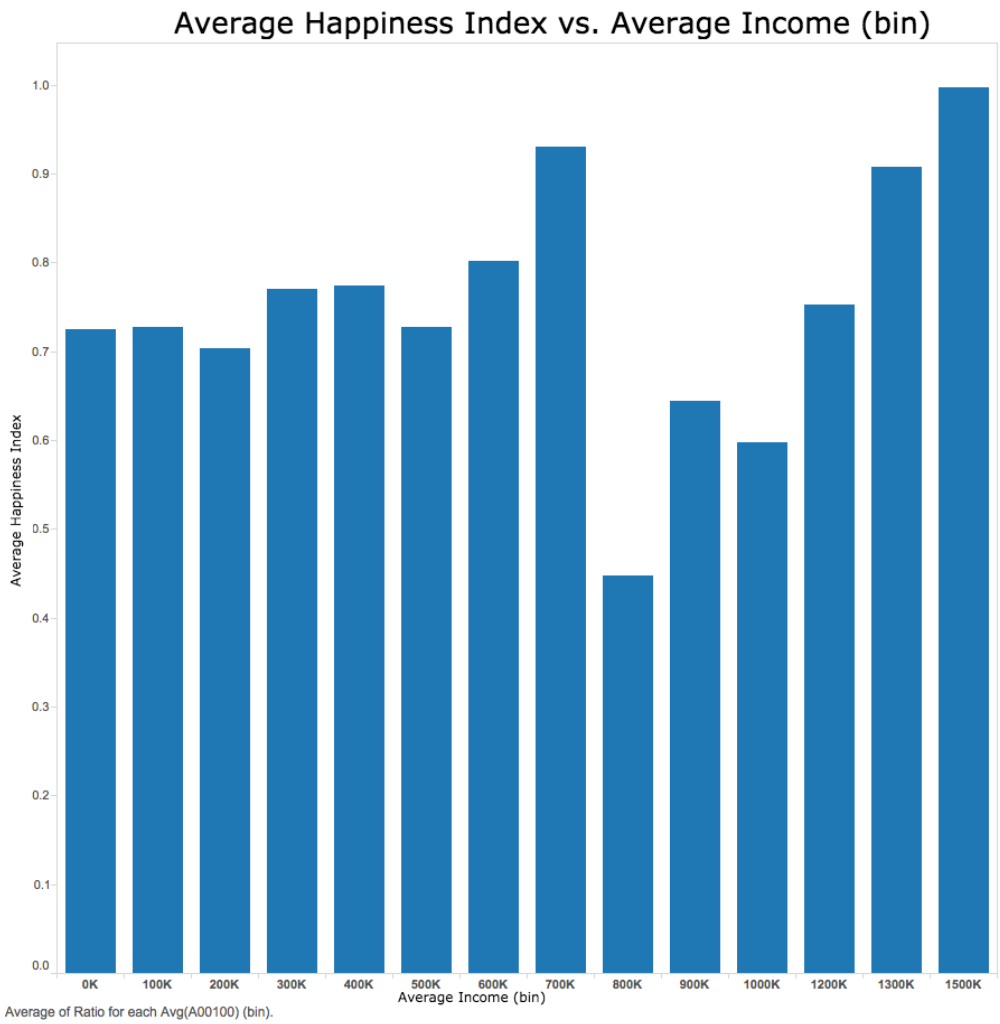

Figure 4: histogram of nationwide happiness by income

Since this graph also shows the same dip at $800K, we do not believe that there is a single area that has such a high cost of living that $800K yearly income isn't sufficient. It is possible that there is some data loss caused by aggregation of income for an entire zip code. For instance, if Warren Buffet moved to an otherwise very low income zip code, their yearly income would spike, but that is unlikely for the same reason: there are many zip codes that exhibit this behavior.

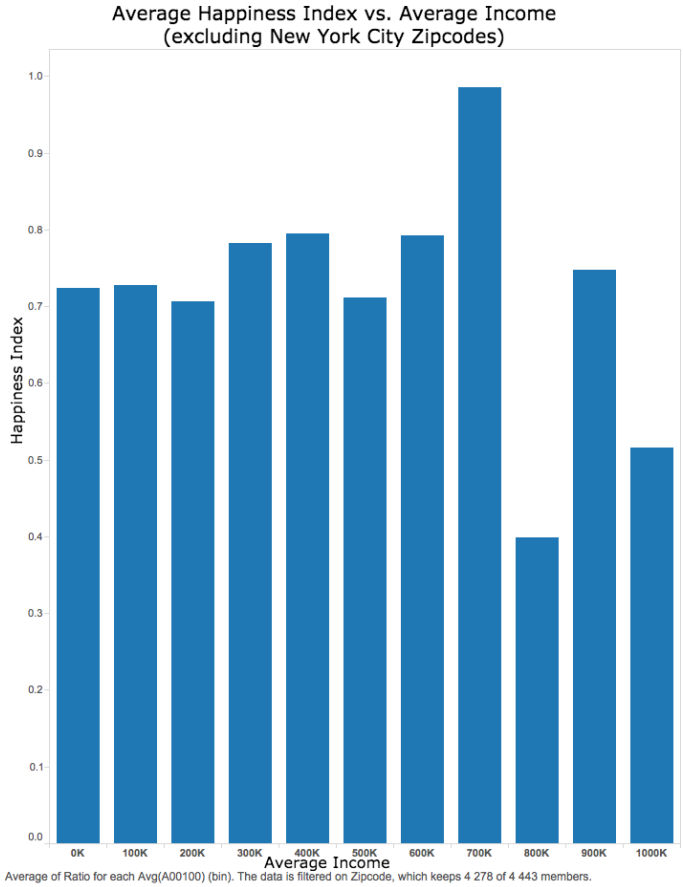

Figure 5: Nationwide happiness by income excluding New York

Conclusion

Although there does seem to be some correlation between income and happiness, the exact nature of it appears complex and non-linear. Much stronger correlations were found between suicide rate and happiness ratio. A paradoxical correlation shows that zip codes of negative sentiment trend toward fewer suicides. The happiness index in conjunction with the income data reveals a stronger trend-line for our suicide rate findings, suggesting that the happiness index alone is not enough to predict social trends and temperament.

Acknowledgements

We would like to thank Dr. Phillip Cannata at the University of Texas at Austin for his mentorship and support of our research as well as the use of his Oracle database and computing cloud.

Our code is publicly available on GitHub. (*8*)

References and Notes

1. Campbell, R. S., and J. W. Pennebaker, *The Secret Life Of Pronouns: Flexibility in Writing Style and Physical Health*, Psychological Science 14.1 (2003), 60–65, Web.

2. Freud, Sigmund. Psychopathology Of Everyday Life 1949. Print.

3. US Atlas `https://github.com/mbostock/us-atlas`

4. Geospatial Data Abstraction Library `https://www.npmjs.com/package/ogr2ogr`

5. Twitter Turns Six `https://blog.twitter.com/2012/twitter-turns-six`

6. Microsoft Depression Research `http://research.microsoft.com/apps/pubs/default.aspx?id=192721`

7. Neal Caren Positive and Negative Word Lists `http://www.unc.edu/~ncaren/haphazard/negative.txt` `http://www.unc.edu/~ncaren/haphazard/positive.txt`

8. Public Github with source code `https://github.com/bolivier/370Research`