Does money make people happy?

Junru Lu, Shijia Gu, Asilayi Bahetibieke, Jingtian Zhou

0. Introduction

Social media nowadays has long been an effective way to express people's feelings, thus gives the city planner a new perspective to do related investigation among citizens. Twitter provides information about citizens' activities, emotions, and their mobility patterns by tweeting with text and emojis. Thus, by combining Twitter data with socio-economic data, we could investigate the relationship between people's emotions and their economic conditions.

1. Literature Review

The development of ICT and big data technology provides many new methods in the city planning area. Naaman (2012) provides a way to rank city based on the intensity of residents' online activities by comparing and analyzing the distribution of Twitter data[1]. Wakamiya (2011) used more than 12 million Twitter data with tags to analyze spatial patterns of people's daily activity [2]. Ljubešić, Nikola and Darja Fišer (2016) did a first attempt at investigating the global distribution of emojis in the paper "A global analysis of emoji usage"[3].

2. Data

We basically used two datasets to conduct the research. The first dataset is the socio-economic data, which includes the mean household income, NYC land use (PLUTO) [4] and 2013-2017

ACS 5-year estimate data [5] that includes basic statistical features. The other dataset is the emotion data, which are acquired by Twitter API [6] to crawl the geo-tagged tweets data posted in NYC area from 2018-10-25 to 2018-11-26. We believe by analyzing the 'emotional words' and emojis in those tweets, we could have a real-time basic NYC emotional map.

3. Methodology

3.1. Data Collection

For the socio-economic data, we directly downloaded the mean household income data from the United States Census Bureau website and two assessment data from NYC PLUTO and ACS data from American FactFinder. For the emotion data, we built a crawler based on Twitter API to get all the tweets published in the NYC area. We've garnered about 424,000 tweets in total where each tweet contains its unique id, the content, the posted time and the posted geolocation.

3.2. Data Preprocessing

3.2.1. Income data

The Mean Household Income Data in NYC is from U.S. Census Bureau, 2012 - 2016 5-Year American Community Survey. Firstly we groupby the data by zipcode. Then we drop redundant columns and keep *Zipcode* column and the *Mean Household Income* column. Finally, we find that there are some NAN values in data. So we clean up data by dropping these NAN values.

3.2.2. PLUTO data

The PLUTO data about land use in New York from Zola are five datasets of five boroughs, so we made a concat of them to get a city-scale dataset. Next, we created a new column about the

assessment of property by subtracting the assessment of land from the total assessment. Then, we transformed the XY coordinates expressed in the New York-Long Island State Plane coordinate system into geographic coordinates for later mergence. Afterwards, we created a GeoDataframe from the csv file with coordinates and aggregate the data from coordinate level into zip level. At last, we filtered out the residential and mixed commercial residential buildings because the census and economy data are collected based on households.

3.2.3. ACS data

The initial ACS data is for the entire state of New York, so we combined the ACS data with NYC zip data to select out the NYC part. After that, 222 zip areas of NYC are selected. Figure 1 shows the statistical properties of the data.

We observe that some columns are full of NAN values. Thus, we need to do further cleanup. We first select all columns that contain values that are not completely NAN, with the minimum of their length as the benchmark, eliminating all rest useless features. After that, the NYC ACS data is filtered again to clear all rows containing NAN. The process removes around 90 columns and 30% rows. However, the zip-level NYC ACS data still contains 459 columns, which is too larger than its observations. So, we use PCA algorithm to reduce dimensionality.

3.2.4. Emotion data

Tweet data were captured within this longitude and latitude range: [-74, -73] and [40, 41]. We use the *sjoin* function in the geoPandas to match the Twitter to the corresponding zip area. However, this is still far from our requirements. Through the time series analysis shown in

Figure 2, we can find that people's tweeting has an obvious circadian pattern. So, we select out the tweets possibly posted in a residential area by limiting their posting time.

After having the appropriate tweet data, we draw on the method of sentiment analysis in natural language processing, which uses the sentiment dictionaries based on large-scale statistics to quantify the positive, negative and neutral sentiment scores of each tweet.

We looked for two sets of sentiment dictionaries, one is SentiWordNet [7] from NLTK package for English words and the other is Emoji Sentiment Ranking for emojis [8]. The constructions of them are shown in Figure 3.

The next thing to do is segmentation. We remove all URLs, punctuation, and stop words in each tweet, then split the tweets into tokens and lemmatize all tokens. Finally, we use the average of the positive scores of each token to get the scores of the sentence. Figure 4 shows the results.

Among four scores, *emoji score*, *text score* and *content score* are calculated by using *split emoji*, *split text* and *split content*, respectively. And *score* equals 50% *emoji score* plus 50% *text score*. The reason that we give emoji and text equal weights is because in some cases that people use diverse languages or social-media-kind way of expression on twitter. Figure 5 shows some cases. Finally, we filter out 70,000 rows with a non-zero *score* as the final available data.

3.3. Time-spatial Analytics (Please refer to Appendix I)

3.4. Correlation Analytics

3.4.1. Correlation

In Figure 9, the correlations between the first group of socio-economic features and four emotion scores show that these features are around 20%~30% positive correlated with our emotion.

3.4.2. Regression

We use the Ridge and Lasso models for regression. Our target value is the *score* column. The entire data set is divided into training and testing sets by 9:1. Table 1 shows the evaluation on both models with optimal α. The model consisting of income, house price and land price has about 10~15% explanatory ability.

3.4.3. Classification

Since the prediction results on regression models are poor, we decide to discretize the target *score*. The result of discretization can be understood as that we divide the emotion score into multiple degrees from weak to strong. The processed data are shown in Figure 10. We use the Logistic models. Table 2 shows the evaluation with the param C=10000.

4. Conclusion

Based on the regression result, wealth and happiness are positively related, but the relationship is not decisive. However, socio-economic data can help generally judge the intensity of a person's positive emotion according to the classification accuracy (0.4797). We can tell that happiness is not dependent on wealth! There are many determinants, which reveal more significance than money, such as sense of safety, love and belonging, esteem and self-actualization that lead people towards happiness. Money has 15% impact on happiness, the rest 85% depends on ourselves.

Reference

1. NAAMAN, M.; ZHANG, A.; BRODY, S.; LOTAN, G.. On the Study of Diurnal Urban Routines on Twitter. International AAAI Conference on Web and Social Media, North America, may.
2. Wakamiya S., Lee R., Sumiya K. (2011) Urban Area Characterization Based on Semantics of Crowd Activities in Twitter. In: Claramunt C., Levashkin S., Bertolotto M. (eds) GeoSpatial Semantics. GeoS 2011.
3. Ljubešić, Nikola, and Darja Fišer. "A global analysis of emoji usage." In *Proceedings of the 10th Web as Corpus Workshop*, pp. 82-89. 2016.
4. https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page.
5. https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.
6. https://developer.twitter.com/en/docs.html.
7. Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." In *Lrec*, vol. 10, no. 2010, pp. 2200-2204. 2010.
8. Novak, Petra Kralj, Jasmina Smailović, Borut Sluban, and Igor Mozetič. "Sentiment of emojis." *PloS one* 10, no. 12 (2015): e0144296.

Appendix I (Time-spatial Analysis)

Income and PLUTO data

Figure 6 shows the choropleth plot of mean household income in NYC area based on zip codes. Downtown Manhattan and downtown Brooklyn have the highest mean household income. Figure 7 shows the choropleth maps of assessed land (left) and property (right) value on the basis of zip codes reveal similar patterns. Manhattan has the highest land and housing prices.

Emotion data

We briefly analyzed the posting number of geo-tagged tweets and the sentiment scores of these tweets from a time-spatial view. Figure 8 shows several time-spatial maps. The left plotting shows people in NYC are more likely to post geo-tagged tweets in mid & low Manhattan, downtown Brooklyn and JFK (Maybe to memory that they are arriving or leaving the city). And a group of four plots in the right side describes the positive scores of these tweets.

Appendix II (Figures and Tables)

	zipcode	GEO.id2	HC01_VC03	HC02_VC03	HC03_VC03	HC04_VC03	HC01_VC04	HC02_VC04	HC03_VC04	HC04_VC04	...	HC03_VC17
count	222.000000	222.000000	222.000000	222.000000	222.000000	0.0	222.000000	222.000000	196.000000	196.000000	...	195.00000
mean	10753.373874	10753.373874	32484.666667	1102.792793	32484.666667	NaN	20618.581081	930.319820	64.403061	2.435204	...	17.68410
std	585.545999	585.545999	23900.798082	602.352291	23900.798082	NaN	15142.590349	524.704641	8.552726	1.799868	...	11.24392
min	10001.000000	10001.000000	0.000000	11.000000	0.000000	NaN	0.000000	11.000000	17.600000	1.000000	...	0.00000
25%	10156.000000	10156.000000	11858.500000	731.250000	11858.500000	NaN	7052.750000	619.250000	59.300000	1.600000	...	9.35000
50%	10474.500000	10474.500000	29858.000000	1185.000000	29858.000000	NaN	18910.500000	946.500000	64.000000	2.000000	...	14.50000
75%	11356.750000	11356.750000	52316.500000	1565.500000	52316.500000	NaN	34519.250000	1331.500000	68.825000	2.600000	...	24.80000
max	11697.000000	11697.000000	86579.000000	2381.000000	86579.000000	NaN	57684.000000	2202.000000	89.700000	18.900000	...	62.50000

Figure 1. Statistical properties of zip-level NYC ACS data

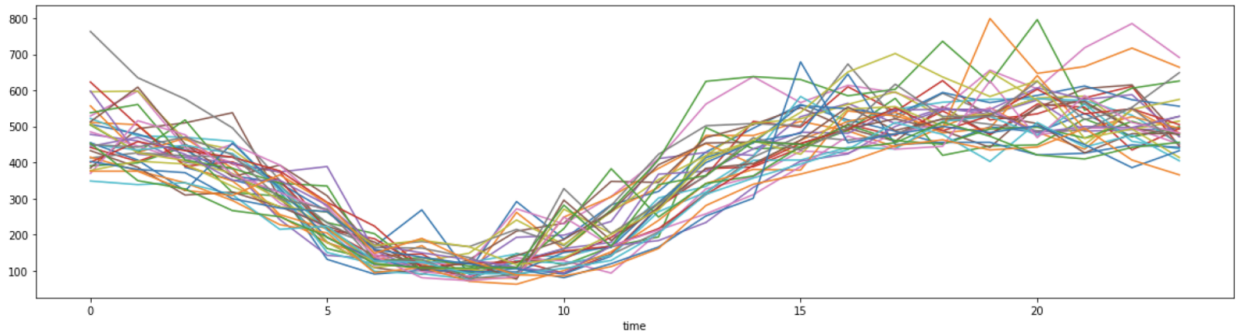

Figure 2: Time series of NYC tweeting in every day from 2018-10-25 to 2018-11-26.

	pos	word_en	word_sp	positive	negative	objective	index	synset	meaning
0	n	neurotropism	NaN	0.125	-0.0	0.875	0	862310	an affinity for neural tissues
1	a	ensorcelled	NaN	0.125	-0.0	0.875	0	865765	under a spell
2	a	bewitched	NaN	0.125	-0.0	0.875	0	865765	under a spell
3	a	mesmerized	NaN	0.125	-0.0	0.875	0	865848	having your attention fixated as though by a s...
4	n	training	NaN	0.125	-0.0	0.875	0	893955	activity leading to skilled behavior

	Emoji	Unicode codepoint	Neg [0...1]	Neut [0...1]	Pos [0...1]	Unicode name
0	😂	0x1f602	0.247	0.285	0.468	face with tears of joy
1	❤	0x2764	0.044	0.166	0.790	heavy black heart
2	♥	0x2665	0.035	0.272	0.693	black heart suit
3	😍	0x1f60d	0.052	0.219	0.729	smiling face with heart-shaped eyes
4	😭	0x1f62d	0.436	0.220	0.343	loudly crying face

Figure 3: SentiWordNet and Emoji Sentiment Ranking

	zipcode	content	coordinates	time	split_content	split_emoji	split_text	emoji_score	text_score	content_score	score
0	11372.0	Midnight Meet & Greet w/ Pastor Boto &...	[-73.8922265, 40.7472744]	2018-11-17 07:32:14	[Midnight, Meet, amp, Greet, Pastor, Boto, amp...	[]	[Midnight, Meet, amp, Greet, Pastor, Boto, amp...	0.00000	0.1250	0.12500	0.062500
0	11372.0	This was so delicious! Authentic shrimp #taco...	[-73.88943573, 40.74729162]	2018-10-27 00:36:55	[This, wa, delicious, Authentic, shrimp, taco,...	[]	[This, wa, delicious, Authentic, shrimp, taco,...	0.00000	0.7500	0.75000	0.375000
0	11372.0	Here with my #team 👹 ❤️ 😉 😊 😘 @ Club Evolution ht...	[-73.88884474, 40.74733265]	2018-11-13 01:43:09	[Here, team, 👹 , ❤, , 😉 , 😊 , 😘 , Club, Evolution]	[👹 , ❤, 😉 , 😊 , 😘]	[Here, team, , Club, Evolution]	0.70275	0.0000	0.70275	0.351375
0	11372.0	Hanging with the girls kit.wonder and ninibeen...	[-73.89127, 40.74749]	2018-11-26 02:19:19	[Hanging, girl, kit, ninibeenie510, jacksondin...	[]	[Hanging, girl, kit, ninibeenie510, jacksondin...	0.00000	0.0625	0.06250	0.031250
0	11372.0	She and I are going to have kittens 😺 , lol. @ ...	[-73.88739, 40.7475]	2018-11-02 01:04:24	[She, I, kitten, 😺 , lol, Amor, karaoke, Bar, a...	[😺]	[She, I, kitten, lol, Amor, karaoke, Bar, amp,...	0.60500	0.1250	0.36500	0.365000

Figure 4: NLP and sentiment quantification

```
sentence = tweets_pd.iloc[200000, 0].lower()
print(sentence)
for word in sentence.replace("#", " ").replace("!", "").strip().split(' '):
    result = word_pd[word_pd.word_en == word]['word_en']
    if len(result) > 0:
        print(word, result.to_dict())
```

```
se eu gosto? nao... eu i ❤️  ny
#ny #nyc #timessquarenyc #timessquare #manhattan #tbt #dancer #choreographer em time… https://t.co/2s
i {29189: 'i'}
manhattan {6352: 'manhattan'}
```

```
sentence = tweets_pd.iloc[100000, 0].lower()
print(sentence)
for word in sentence.replace("#", " ").replace("!", "").strip().split(' '):
    result = word_pd[word_pd.word_en == word]['word_en']
    if len(result) > 0:
        print(word, result.to_dict())
```

```
🗽 ⛲ 🌉 🌇 #brooklynbridge #thanksgod🙏 #kiraznewyorkta @ brooklyn, new york https://t.co/cq4fokdzzj
new {13032: 'new', 7723: 'new', 24124: 'new', 38493: 'new', 23303: 'new'}
```

```
S E C O N D P A R K D A Y ❤️ 🌟 🎉 🎠 🎡 🥂
...vamos brincar vamos brincar vamos brincar... !😜 ❤️ 😉 😎 🇺🇸 🙏 👍🏻 🎉
.
.
#disneyland… https://t.co/rUEplldcyT
```

Figure 5: Cases that emojis have rich content

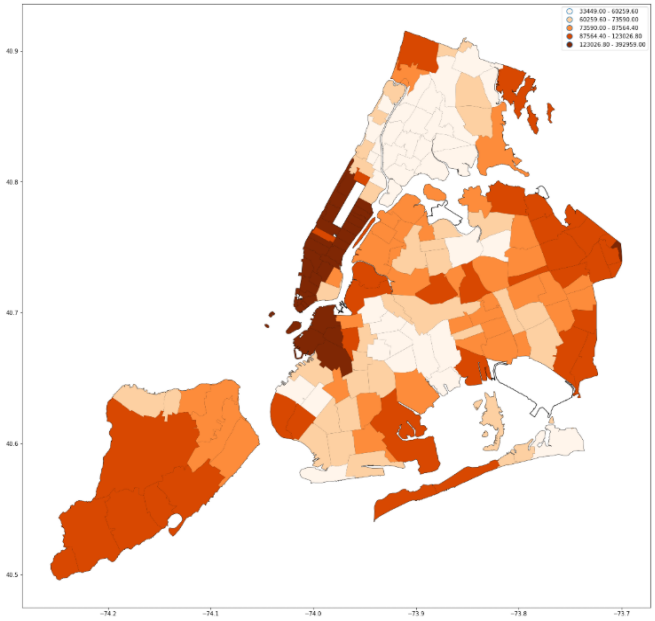

Figure 6: The mean household income based on NYC zips for Fiscal Year 2016

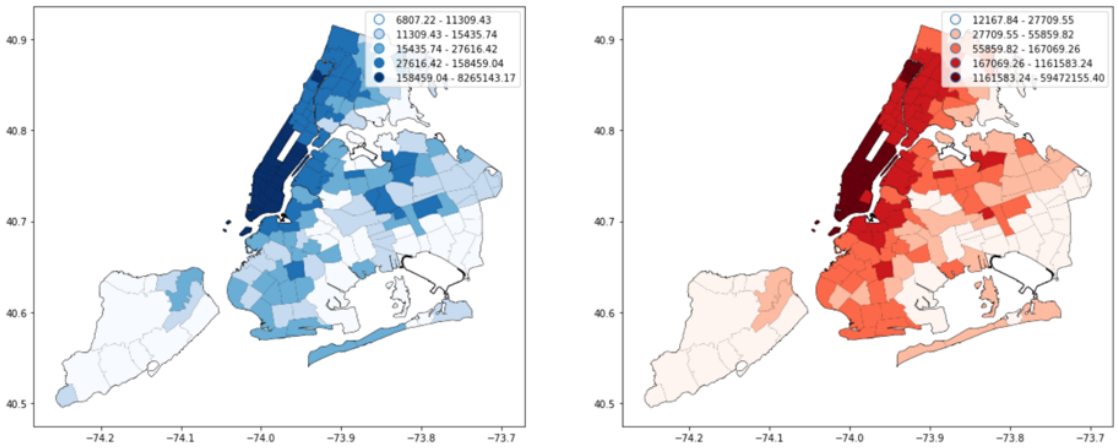

Figure 7: The tentative assessed land (left) and property(right) value for Fiscal Year 2018

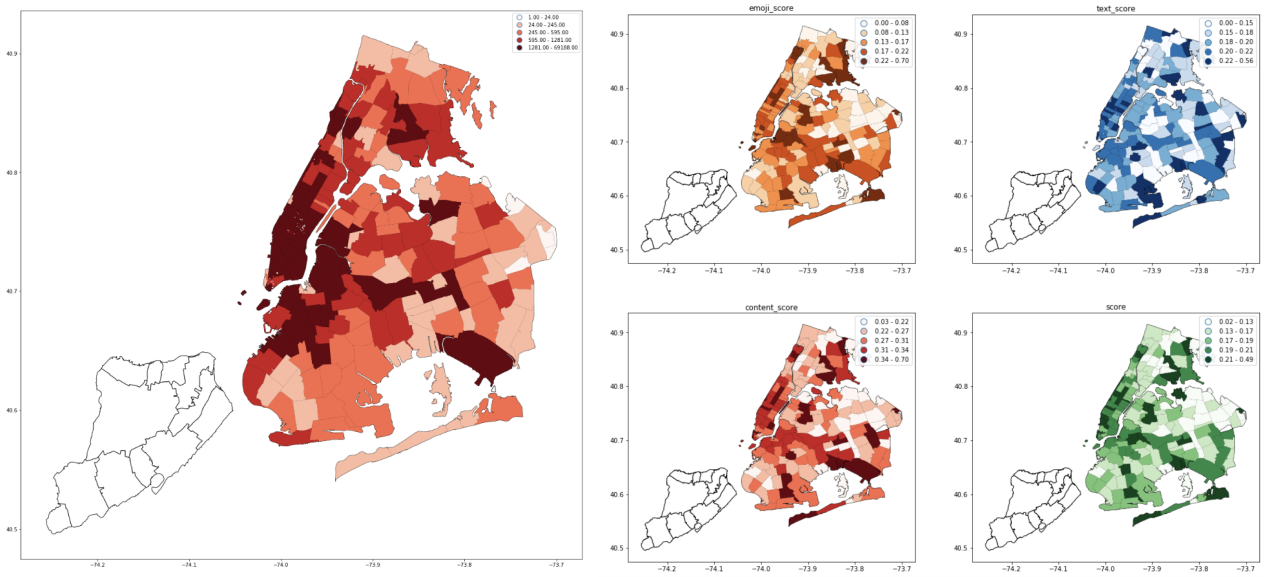

Figure 8: The average amount of geo-tagged tweets and their sentiment scores from 2018-10-25 to 2018-11-26

	AssessLand	AssessProperty	mean_income	content_score	text_score	emoji_score	score
AssessLand	1.000000	0.893704	0.569888	0.355519	0.280824	0.173384	0.294627
AssessProperty	0.893704	1.000000	0.559398	0.386493	0.214809	0.233768	0.312812
mean_income	0.569888	0.559398	1.000000	0.373503	0.372747	0.210275	0.373896
content_score	0.355519	0.386493	0.373503	1.000000	0.543722	0.755687	0.933552
text_score	0.280824	0.214809	0.372747	0.543722	1.000000	-0.014663	0.502686
emoji_score	0.173384	0.233768	0.210275	0.755687	-0.014663	1.000000	0.857005
score	0.294627	0.312812	0.373896	0.933552	0.502686	0.857005	1.000000

Figure 9: Correlation between the 1st group of socio-economic features and target values

	AssessLand	AssessProperty	mean_income	content_score	text_score	emoji_score	score	content_score_dis	text_score_dis	emoji_score_dis	score_dis
0	0.635338	0.531722	1.222160	0.321180	0.202377	0.187339	0.194858	7	5	4	4
1	-0.053623	-0.102250	-0.564562	0.302879	0.189755	0.198565	0.194160	7	4	4	4
2	0.066027	0.006609	1.385603	0.321930	0.218046	0.177017	0.197532	7	5	4	4
3	7.584219	4.220044	1.775342	0.312889	0.224669	0.166850	0.195759	7	5	4	4
4	1.694808	1.411083	1.432299	0.277121	0.185994	0.146513	0.166254	6	4	3	4

Figure 10: Discreted positive emotion scores

Table 1. R-squared of Ridge and Lasso models on two groups of features.

	Group 1 (Income, land assessment and property assessment)	Group 2 (2013-2017 ACS 5-year estimate)
Ridge / Train	0.1573	0.1650
Ridge / Test	0.1241	0.0924
Lasso / Train	0.1287	0.1161
Lasso / Test	0.1007	0.0591

Table 2. Accuracy of Logistic models on two groups of features.

	Group 1 (Income, land assessment and property assessment)	Group 2 (2013-2017 ACS 5-year estimate)
Train	0.4797	0.5563
Test	0.3529	0.1875

Appendix III (Shortage)

In this project, there are at least 5 inevitable data errors as follows:

1. Not everyone will express emotions on social media, or express positive emotions. There is a remarkable age limitation because young people are more likely to express emotions on social media.

2. We tried to further remove tweets that posted in non-working and residential areas by *sjoin* tweets data with PLUTO data, but the processing time is too long to complete.

3. Some tweets were posted by visitors and official accounts other than residents.

4. There is an error in the process of sentiment quantification because we did not find a particularly detailed sentimental dictionary for emojis.

5. After discretizing the emotional scores, we use classification instead of regression, which neglects the comparability between discrete labels. For example, 5 should indicate a stronger emotion than 4.

Appendix IV (Contribution)

Junru Lu: Twitter data acquisition and preprocessing; Correlation analysis

Shijia Gu: PLUTO data acquisition and preprocessing; Time-spatial analysis

Asilayi Bahetibieke: Income data acquisition and preprocessing; Time-spatial analysis

Jingtian Zhou: ACS data acquisition and preprocessing; Correlation analysis