Weather impacts expressed sentiment of posts on social media

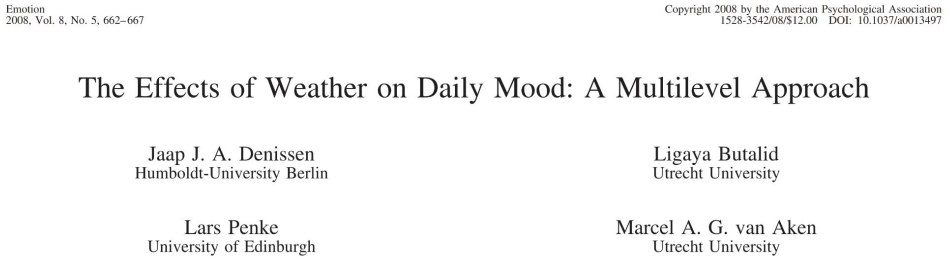
Xi Chen, Rongfan Liu

Getting started...

Good weather and good spirit!

Some theories behind...

Denissen et al. (2008) point out that good weather has a positive effect on people's mood while bad weather has a negative effect.



Emotion
2008, Vol. 8, No. 5, 662–667

1528-3542/08/$12.00 DOI: 10.1037/a0013497

The Effects of Weather on Daily Mood: A Multilevel Approach

Jaap J. A. Denissen
Humboldt-University Berlin

Ligaya Butalid
Utrecht University

Lars Penke
University of Edinburgh

Marcel A. G. van Aken
Utrecht University

We want to know to what extent can weather impact our mood

Some limitations on questionaires (used for collecting emotion data in this study)

With the help of Data Science! Maybe we can do more!

Outline

Overview → Getting Data

→ Train and evaluate model → Reflection

Overview

Goal: Setting up a quantified model of the weather and people’s emotion

Data: 1. Weather data 2. people’s emotion data (posts on weibo)

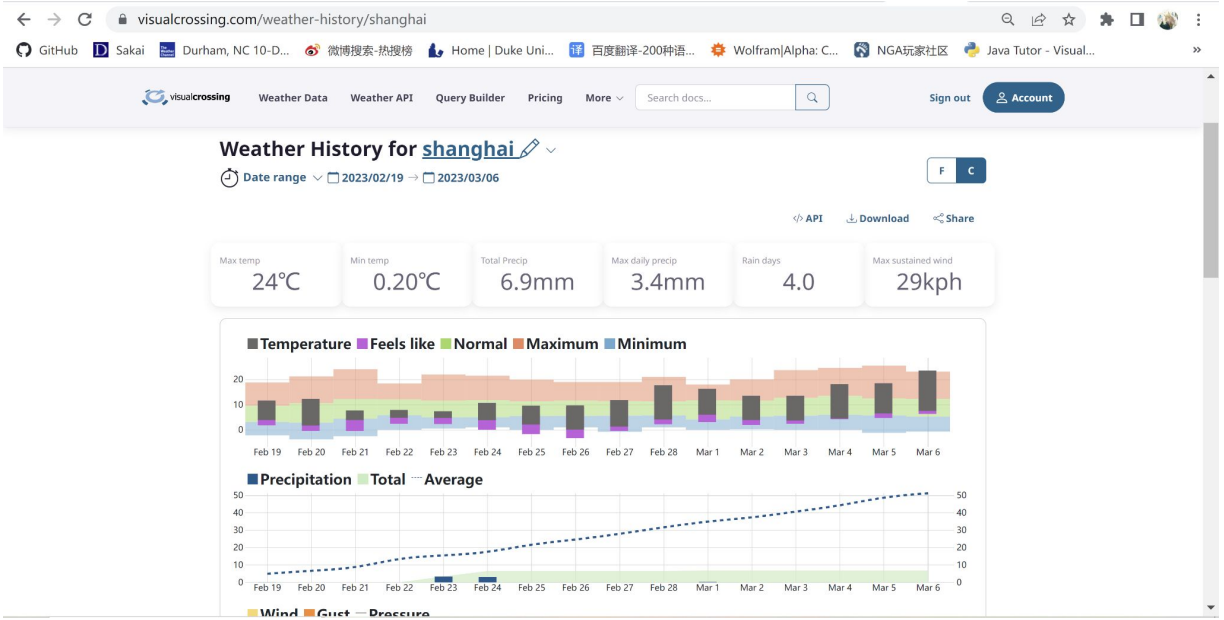
Expected output: An numerical model which can describe and predict the people’s emotion based on weather

Getting the data-Weather Data

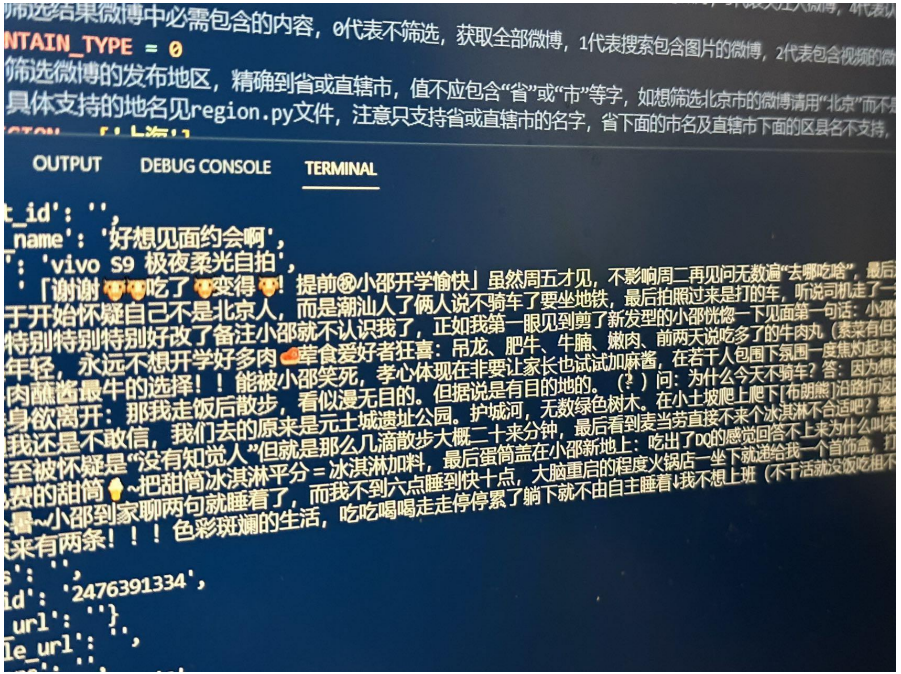
Visual Crossing— An API to download history weather data

We download the weather data of Shanghai from 2022-03-05 to 2023-03-05

Weather data includes max/min/average temperature; humidity; UV index ect.



Getting the data- Posts on social media



Use the spider to get the posts containing the keyword "上海 天气" (Shanghai Weather) in the same time period as weather data on 微博 (Weibo).

Text and Time are what we are looking for

We get around 40,000 posts in total.

Processing the data—NLP

Using Stanza— a python natural language processing package to sense the emotion of the text

Positive 😀 Sentence → 2

Neural 😐 Sentence → 1

Negative 😭 Sentence → 0

If multiple sentences in a post?
Average it!

Processing the data—Summary and Combination

Post data: After NLP, we get a list of number (Emotion Index) with its corresponding posting time.

We need to combine the post data and the weather data to make a completed dataset.

“Look up” the weather data sheet with the Posting time as a key!

and combine the weather data with the Emotion Index

Data Description

33 columns each with different attributes, with 39804 rows

We neglect some of value including location, icon, name and duplicate values

Catogory value are quantified by one hot encoding

The final attributes for X has 31 features.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	name	datetime	tempmax	tempmin	temp	feelslikem	feelslikem	feelslike	dew	humidity	precip	precipprol	precipcov	precipty
2	shanghai	2022/3/5	15.6	7.7	11.6	15.6	5.5	11.1	2.1	58.6	0	0	0	
3	shanghai	2022/3/6	12.7	7.6	9.1	12.7	5.4	7.8	-2.5	45.1	0.7	100	8.33	rain
4	shanghai	2022/3/7	15.3	6.9	10.4	15.3	6	9.9	1.3	55.2	0.7	100	12.5	rain
5	shanghai	2022/3/8	19.7	6.6	12.7	19.7	6.2	12.6	4.3	59.5	0	0	0	
6	shanghai	2022/3/9	20.4	8.3	14.1	20.4	7.5	13.8	5.3	57.5	0	0	0	
7	shanghai	2022/3/10	21.3	11.6	15.9	21.3	11.6	15.9	9.5	66.8	0	0	0	
8	shanghai	2022/3/11	25.5	13.2	18.8	25.5	13.2	18.8	11.9	67.5	0	0	0	
9	shanghai	2022/3/12	25.8	13.8	19.4	25.8	13.8	19.4	14.6	75.9	0	0	0	
10	shanghai	2022/3/13	27.8	16.6	21.3	28	16.6	21.3	15.1	69.2	0.7	100	4.17	rain
11	shanghai	2022/3/14	29	13.3	18.5	29.5	13.3	18.5	13.1	72.3	0.3	100	4.17	rain
12	shanghai	2022/3/15	21.5	10.5	15.5	21.5	10.5	15.5	7.4	61.1	0	0	0	
13	shanghai	2022/3/16	27.2	13.1	19.8	27.5	13.1	19.8	13.7	69.7	0	0	0	
14	shanghai	2022/3/17	19.3	10.2	15.4	19.3	10.2	15.4	14.9	96.5	12.9	100	16.67	rain
15	shanghai	2022/3/18	11.1	7.2	9.3	11.1	3.5	7.4	7.2	87.5	1	100	8.33	rain
16	shanghai	2022/3/19	13.4	7.5	10.2	13.4	4.3	9	6	75.3	0.1	100	4.17	rain
17	shanghai	2022/3/20	8.2	7	7.8	6.6	3.8	5.5	5.1	84.1	23.1	100	16.67	rain
18	shanghai	2022/3/21	9	7.8	8.2	6.3	4	5.1	7.4	95.1	56.4	100	29.17	rain
19	shanghai	2022/3/22	10.4	7.6	9.1	10.4	4.4	7.3	4.8	75.6	1.9	100	4.17	rain
20	shanghai	2022/3/23	14.5	7.2	10.3	14.5	4.3	9.3	2.1	58.9	0.3	100	8.33	rain
21	shanghai	2022/3/24	18	6	12.6	18	4.5	12.2	5.6	64	0	0	0	
22	shanghai	2022/3/25	19	14.8	17.1	19	14.8	17.1	15.8	91.7	25.1	100	16.67	rain
23	shanghai	2022/3/26	16.9	12.5	14.7	16.9	12.5	14.7	7.8	64.8	0.1	100	4.17	rain
24	shanghai	2022/3/27	18.5	10.8	14	18.5	10.8	14	4	52.9	0	0	0	
25	shanghai	2022/3/28	14.7	9.5	11.6	14.7	8	11.4	3.1	56.5	0	0	0	
26	shanghai	2022/3/29	19.4	6.9	13.5	19.4	5.9	13.2	5.1	59	0	0	0	
27	shanghai	2022/3/30	24.6	13.4	17.8	24.6	13.4	17.8	12.2	71.4	0	0	0	rain
28	shanghai	2022/3/31	13.8	9.2	12	13.8	6	11.7	7.9	76.4	1.1	100	4.17	rain
29	shanghai	2022/4/1	13.8	7.3	10.5	13.8	4.2	9.1	2.1	57	0	0	0	
30	shanghai	2022/4/2	16.1	5.9	10.8	16.1	3.8	10	1.5	57	0	0	0	

shanghai_weather

Data Description

The sentiment analysis score is calculated using the Stanza library, each sentence has its own sentiment value. The final result is the average of sentiment values from different sentences.

The sentiment data then merge with the weather data regarding the time column.

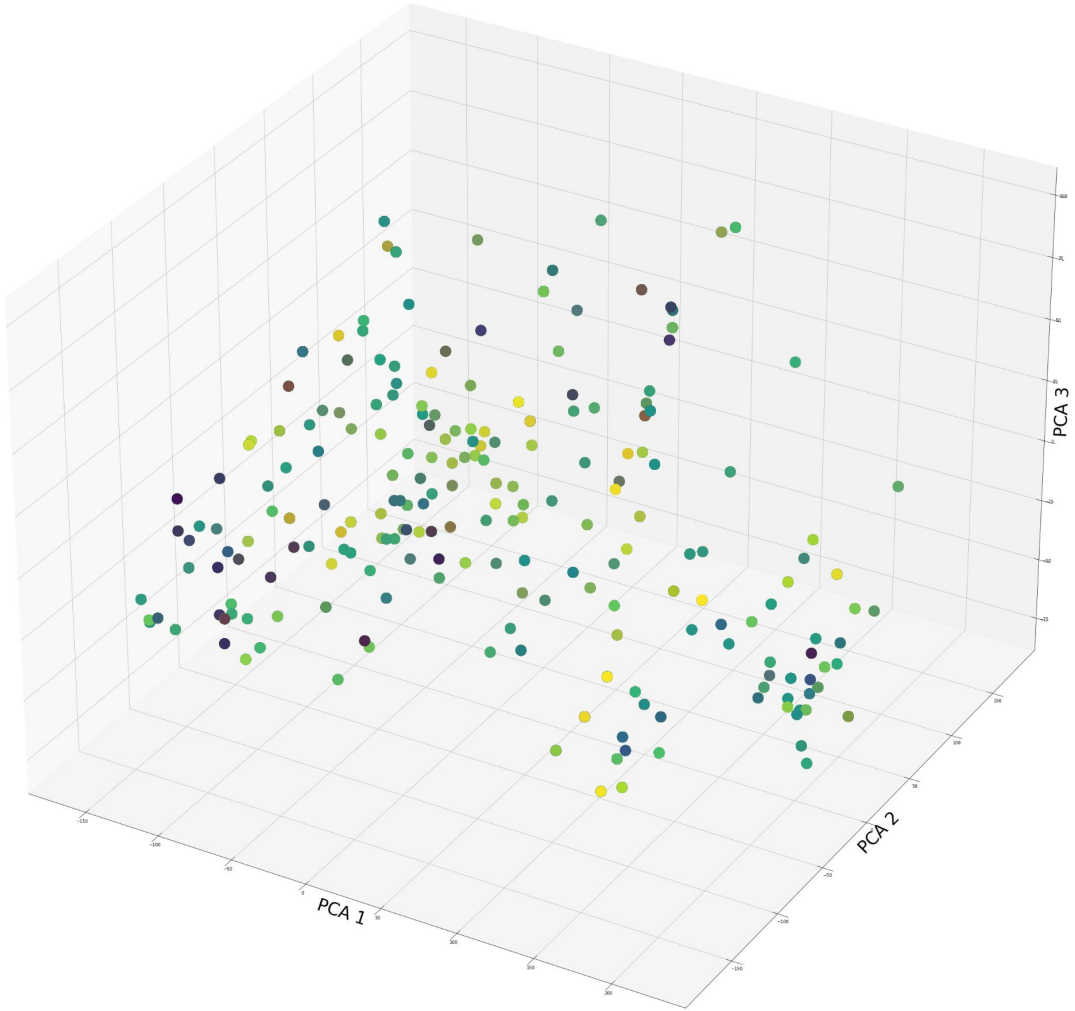
The sentiment value are labels (Y).

	text	time	sentimen
0	喜欢色戒和罗曼蒂克消亡史这两部，看看过去的上海，不仅感慨经典已成过去，罗曼蒂克消亡矣……最近...	2022-08-08	2.000000
1	【高温橙色预警！明天，#陕西湖北重庆等局地将达40度#】预计9日，陕西、河南、安徽、江苏、上...	2022-08-08	0.666667
2	现在逛商场好幸福，因为哈尔滨的天气现在已经要穿长袖长裤了，所以短袖T恤连衣裙都会打折，但是回...	2022-08-10	2.000000
3	上海这BT的天气，没有一丝丝风，即使到半夜还汗流浃背，要知道我是个不怎么出汗的人呀	2022-08-09	0.000000
4	跑团上个月就已经成立三周年了，由于疫情的影响，想搞个活动庆祝一下也没法弄。总算等到上海疫情清...	2022-08-08	1.000000
...	...	...	...
39799	晚饭之后一顿压马路现在终于躺在床上挑这首歌作为今天的总结上海现在的天气太适合白天骑车晚上压马路了	2022-09-12	2.000000
39800	#市民盲目放生致苏州河现大量死鱼#近日，上海苏州河上频现大量死鱼。据上海市市容环境卫生水上管...	2022-08-25	1.000000
39801	天气凉得好快呀睡觉盖夏凉被已经开始不暖和了	2022-08-23	0.000000
39802	#上海天气#今天好凉快啊我好开心	2022-08-24	2.000000
39803	因为天气不好	2022-09-13	0.000000

39804 rows × 3 columns

Data Visualization

The data is reduced to 3 components by principle component analysis. The color represents the sentiment value when darker green means positive and lighter green means negative.



Data Preprocessing

1. Data normalization

   Standard Scalar and MinMax Scalar

2. Delete weibo duplicate text for emotion prediction, but keep duplicate text for text analysis.

   Out of almost 40,000 blogs, 5,000 are duplicate

3. Form value X: weather data and Y: sentiment of blogs

Setting up the neural network

Using 1D convolution network, maxpooling, dropout ,and batch normalization

!!The model not improving

```
636/636 [==============================] - 6s 5ms/step - loss: 0.2778 - accuracy: 0.1991
Epoch 2/50
636/636 [==============================] - 4s 7ms/step - loss: 0.2778 - accuracy: 0.1989
Epoch 3/50
636/636 [==============================] - 3s 5ms/step - loss: 0.2777 - accuracy: 0.1990
Epoch 4/50
636/636 [==============================] - 3s 5ms/step - loss: 0.2774 - accuracy: 0.1993
Epoch 5/50
636/636 [==============================] - 3s 5ms/step - loss: 0.2784 - accuracy: 0.1988
Epoch 6/50
636/636 [==============================] - 4s 6ms/step - loss: 0.2770 - accuracy: 0.1995
Epoch 7/50
636/636 [==============================] - 3s 5ms/step - loss: 0.2780 - accuracy: 0.1985
Epoch 8/50
636/636 [==============================] - 4s 6ms/step - loss: 0.2782 - accuracy: 0.1985
Epoch 9/50
636/636 [==============================] - 4s 6ms/step - loss: 0.2772 - accuracy: 0.1996
Epoch 10/50
636/636 [==============================] - 4s 6ms/step - loss: 0.2779 - accuracy: 0.1989
Epoch 11/50
636/636 [==============================] - 3s 5ms/step - loss: 0.2776 - accuracy: 0.1994
Epoch 12/50
636/636 [==============================] - 3s 5ms/step - loss: 0.2774 - accuracy: 0.1996
Epoch 13/50
636/636 [==============================] - 4s 7ms/step - loss: 0.2778 - accuracy: 0.1977
Epoch 14/50
636/636 [==============================] - 3s 5ms/step - loss: 0.2773 - accuracy: 0.1996
Epoch 15/50
636/636 [==============================] - 3s 5ms/step - loss: 0.2782 - accuracy: 0.1985
Epoch 16/50
```

Setting up the regression model

Several regression model were experimented, but the model could not predict any information.

```
from sklearn.linear_model import LinearRegression
reg = LinearRegression().fit(xx[:1000], yy[:1000])
reg.score(xx[1000:1100],yy[1000:1100])
```

```
-0.07011898259830796
```

```
from sklearn import tree
clf = tree.DecisionTreeRegressor()
clf=clf.fit(xx[:1000], yy[:1000])
clf.score(xx[1000:1100],yy[1000:1100])
```

```
-0.3595906224172809
```

Most of the regression model has a negative accuracy

```
from sklearn import svm
regr = svm.SVR()
regr=regr.fit(xx[:1000], yy[:1000])
regr.score(xx[1000:1100],yy[1000:1100])
```

```
-0.06851895366232941
```

Reduce duplication and insignificant information

In the weather data, multiple columns values are correlated,

“max temp”, “min temp”, “average temp”

“snow”, “snow depth”

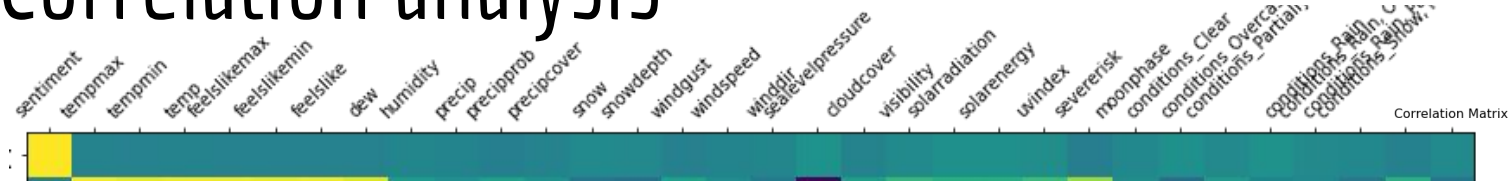
“precip”, “precipprob”

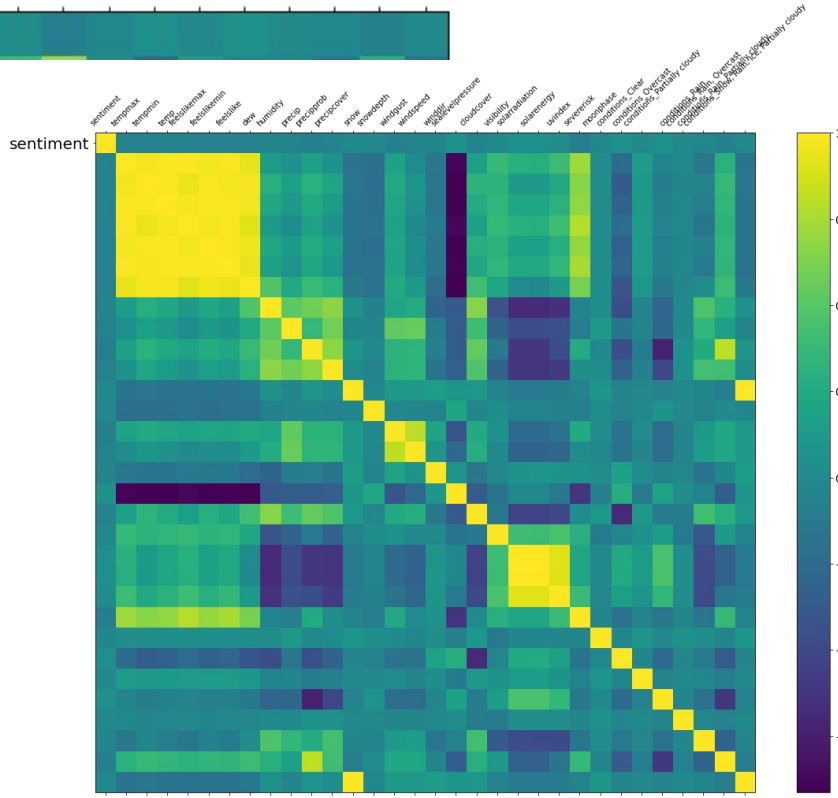
Only select the distinctive features,

Still model not improving

Model error or Data error?

Correlation analysis





After calculating the correlation matrix, it is shown that the sentiment value of the text has almost zero correlation

Top 10 frequent word in blog text

('上海', 55446)('影响', 12796)

('天气', 48813)('地区', 10937)

('今天', 19644)('台风', 9760)

('气温', 16636)('最高', 9588)

('高温', 14683)('视频', 8995)

Possbile reasons behind...

1. The correlations between weather and emotion may not be as strong as we thought: Denissen et al. (2008) also points out that the relationship is kind of weak: Only Temperature can be seen a significant impact ($p>0.01$)

2. Some noise in post data: When we grabbing data from Weibo, we spotted some posts are posted by users who are basically writing diaries, but others are post by social media agencies. These post cannot reflect the actual feeling of human and should not be considered in this analysis. However, we are not able to delete them since we have a large database and the noise posts have no uniform pattern

Reference

The Effects of Weather on Daily Mood: A Multilevel Approach-http://larspenke.eu/pdfs/Denissen_Butalid_Penke_van_Aken_2008_-_Weather_and_mood.pdf

Stanza-https://stanfordnlp.github.io/stanza/

Visual crossing-https://www.visualcrossing.com/

Weibo search-https://github.com/dataabc/weibo-search

Thanks for watching!