Tweet-based Subjective Happiness Index

@ Asia-Pacific Stats Cafe Series
26 July 2021

Siti Mariyah & Asita Sekar Asri

Background

Learning from Survey

✓ Happiness Level Measurement Survey in 2014 and 2017

✓ Composite of Life Satisfaction, Affect and Eudaemonia

✓ Score scales 0-100, national estimate

✓ No longer performed

Social Media Use

❖ 15,1 million of Twitter Users are Indonesian people (April, 2021)

❖ Twitter provides Streaming API with different access grant. It provides 70 variables such as tweet text, user profile, location, and timestamp, etc.

❖ Real time and high velocity

Main Literature

Dodd's et al. (2001) "Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter":
✓ content of tweets does appear to reflect people's current circumstances;
✓ Develop laBMT 1.0 that contains word list and its average happiness representing a sad to happy spectrum

Objective

❋ Find preliminary pattern of tweets and happiness from a set of samples
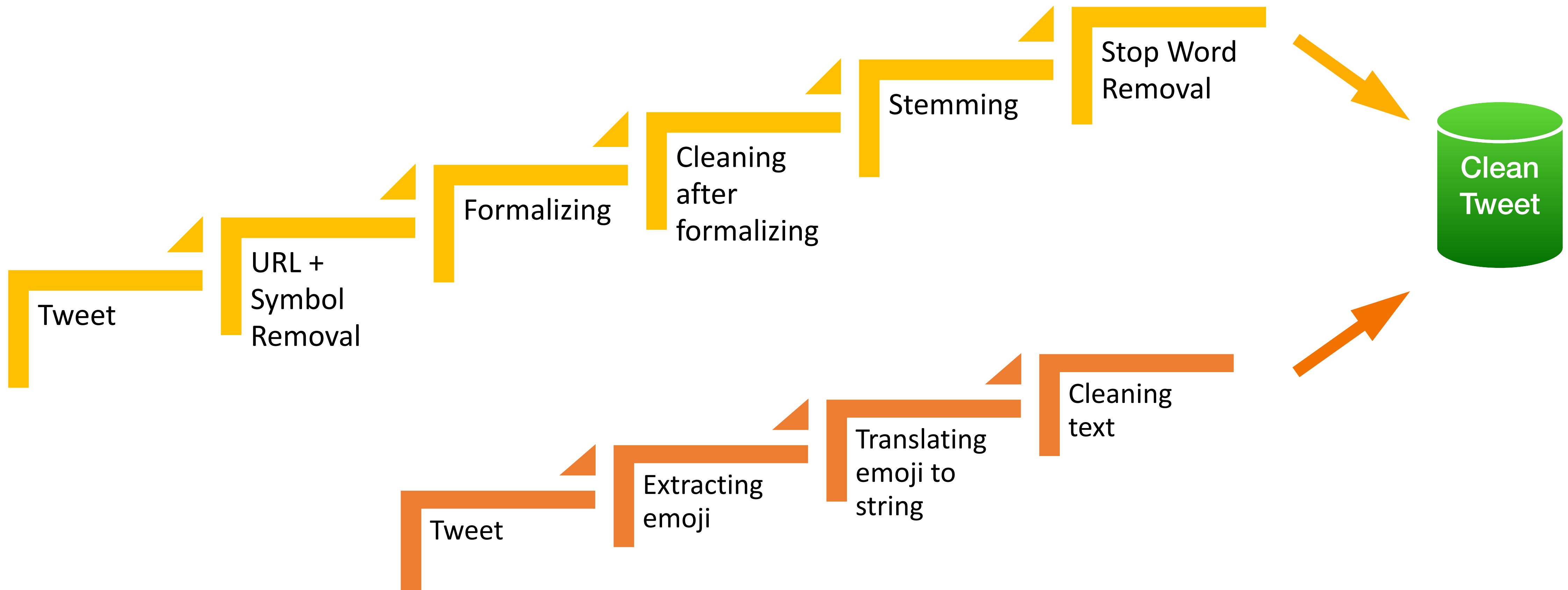
❋ Feasibility study:

◆ complementary dataset

◆ method

Data Collection

Methodology

☀ Apply for Twitter Streaming API access

☀ Set the bounding box to ensure tweets come from Indonesia

☀ Streaming from April to July 2018

☀ Languages: Bahasa Indonesia, Melayu, and English

☀ Tweets samples given by Twitter

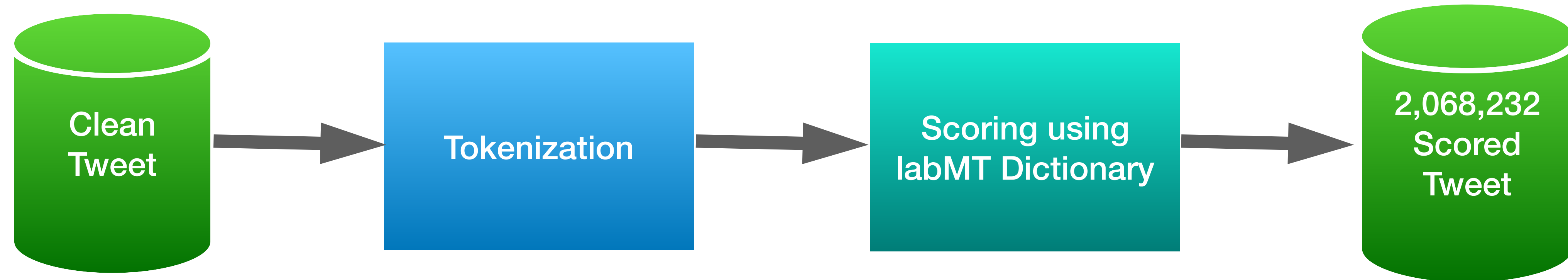# Pre-processing

Methodology

Tweet → URL + Symbol Removal → Formalizing → Cleaning after formalizing → Stemming → Stop Word Removal → **Clean Tweet**

Tweet → Extracting emoji → Translating emoji to string → Cleaning text → **Clean Tweet**

Processing

Methodology

The scoring scales from 1 to 9

The higher scale, the happier the word

$$h_{avg}(T) = \frac{\Sigma_{i=1}^{N} h_{avg}(w_i) f_i}{\Sigma_{i=1}^{N} f_i} = \Sigma_{i=1}^{N} h_{avg}(w_i) p_i$$

Where

T : set of N words in each region

f_i : frequency of i-th word

$h_{avg}(w_i)$: average happiness of i-th word

p_i : normalized frequency

Evaluation

Methodology

⁂ Performed two treatments for evaluation using 100,00 tweets

⁂ One dataset is scored by human (manual labelling) and labMT

⁂ Someone scores 9 to a tweet if it shows the happiest sentiment, 1 to saddest sentiment, and 0 to unknown tweets

⁂ Compared scores from human (manual labelling) and and LabMT

labMT

- Is a software that uses the language assessment by Mechanical Turk (MT) word list to score the happiness of a corpus, provided by Andy Reagan.

- The labMT word list was created by combining the 5000 words most frequently appearing in four sources: Twitter, the New York Times, Google Books, and music lyrics, and then scoring the words for sentiment on Amazon's Mechanical Turk.

	word	happiness_rank	happiness_average	happiness_standard_deviation	twitter_rank	google_rank	nyt_rank	lyrics_rank
1	laughter	1	8.50	0.9313	3600	NA	NA	1728
2	happiness	2	8.44	0.9723	1853	2458	NA	1230
3	love	3	8.42	1.1082	25	317	328	23
4	happy	4	8.30	0.9949	65	1372	1313	375
5	laughed	5	8.26	1.1572	3334	3542	NA	2332
6	laugh	6	8.22	1.3746	1002	3998	4488	647
7	laughing	7	8.20	1.1066	1579	NA	NA	1122
8	excellent	8	8.18	1.1008	1496	1756	3155	NA
9	laughs	9	8.18	1.1551	3554	NA	NA	2856
10	joy	10	8.16	1.0568	988	2336	2723	809

Database

Tweets Distribution in Each Province

Kalimantan Utara
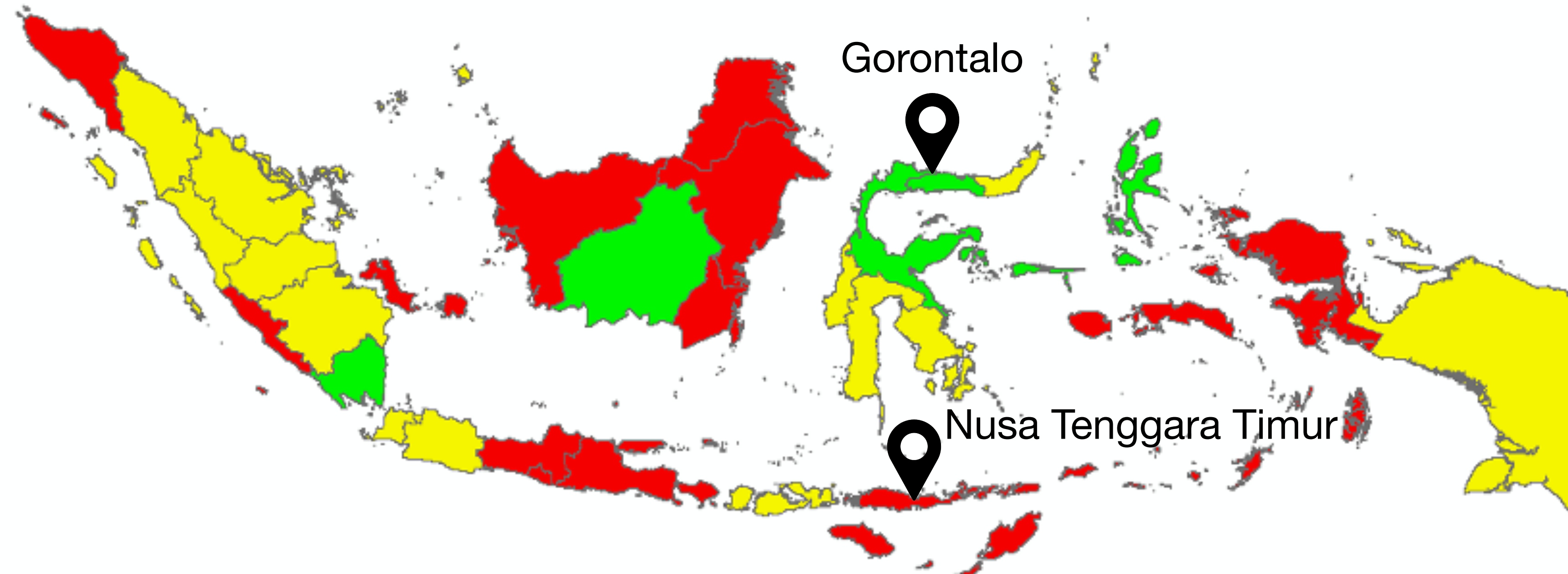
DKI Jakarta &
Jawa Barat
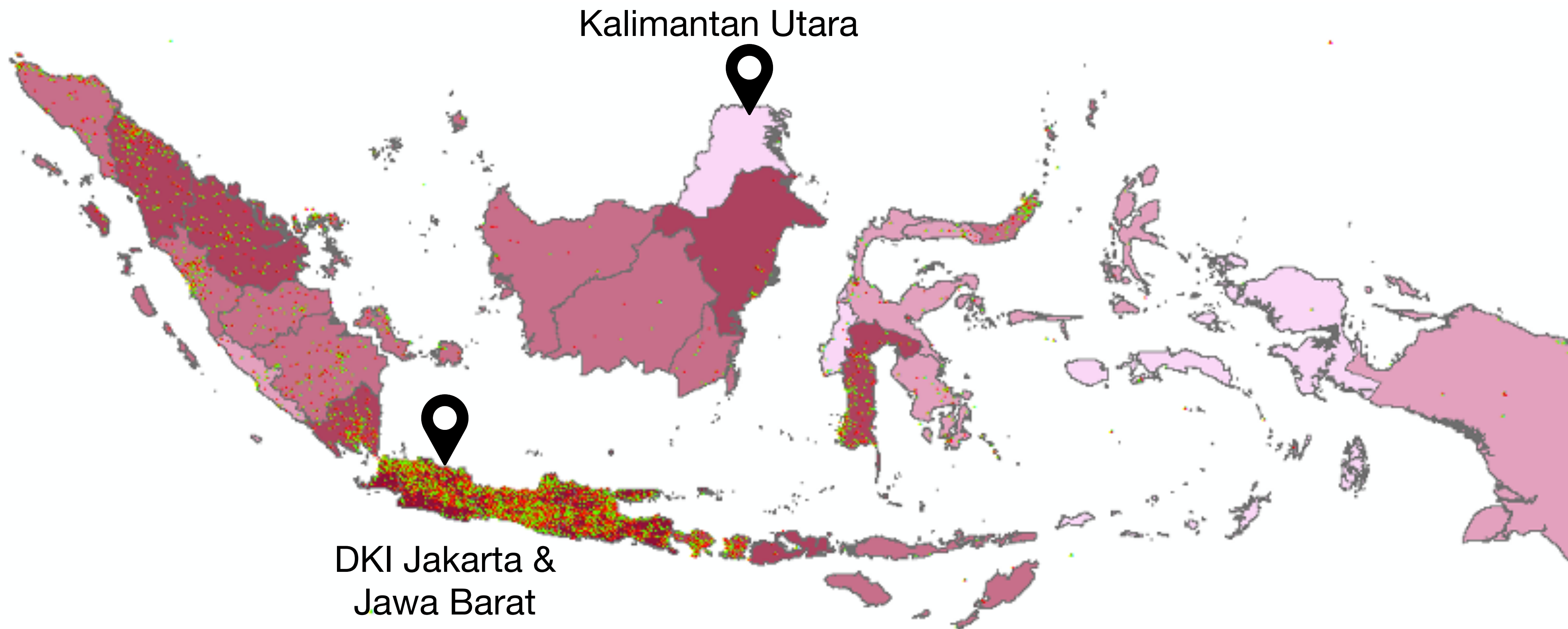
❊ The darker the
colour, the more
tweets posted

❊ The green dots
show happy/positive
sentiment; the red
one means sadness/
negative sentiment

❊ Java contributes
more than 70 per
cent of total tweets

❊ Jawa Barat and
DKI Jakarta as the
two top regions with
the most tweets
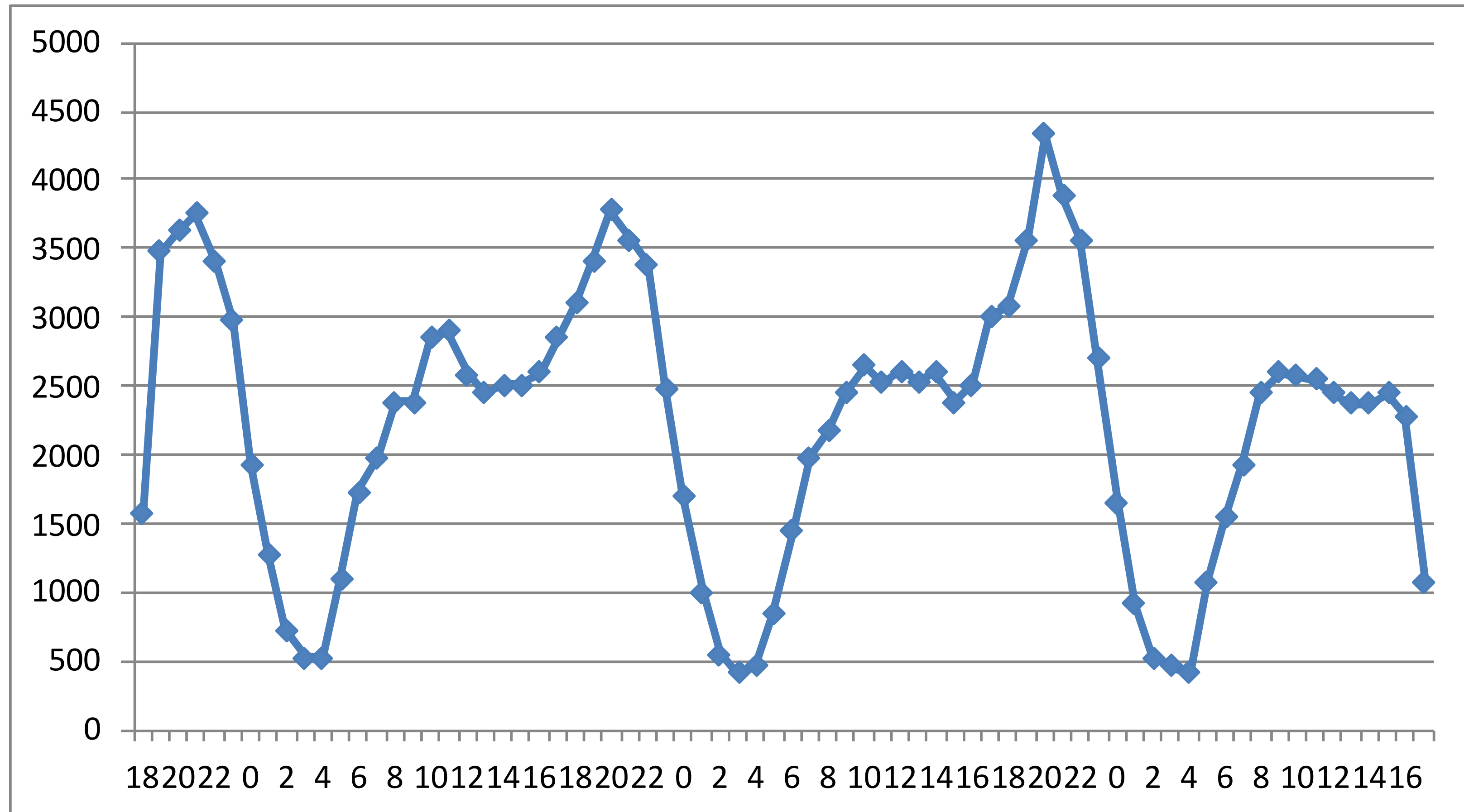
Subjective Happiness Index in Each Province

Gorontalo

Nusa Tenggara Timur

❀ Indonesia's average subjective happiness index is 5.59, with the lowest index is 1.78, and the highest is 8.26.

❀ The centroid of each cluster is 5.56, 5.61, and 5.69. Red is for provinces with the lowest subjective happiness index, and green is for the highest index.

❀ Gorontalo is a province with the highest subjective happiness index, followed by Kalimantan Tengah and Maluku Utara, and Nusa Tenggara Timur is a province with the lowest index.

❀ Coincide with the streaming time, there were several cultural festivals in Gorontalo and found that most of Gorontalo people's tweets were written in a positive tone. While in Nusa Tenggara Timur, most of the tweets were about complaints of jobs, politics, life, and others.

❀ DKI Jakarta (capital of Indonesia) is in the 14th position with a subjective happiness index of 5.6. There was a balanced number of positive and negative tweets.

❀ Compared to the poverty rate in March and September 2018, the subjective happiness index seems contrary to the poverty rate. For example, NTT is the third province with the highest poverty rate, and Kalimantan Tengah is the fourth province with the lowest poverty rate.

Subjective Happiness Index in Each District

Cilegon

Lombok Timur
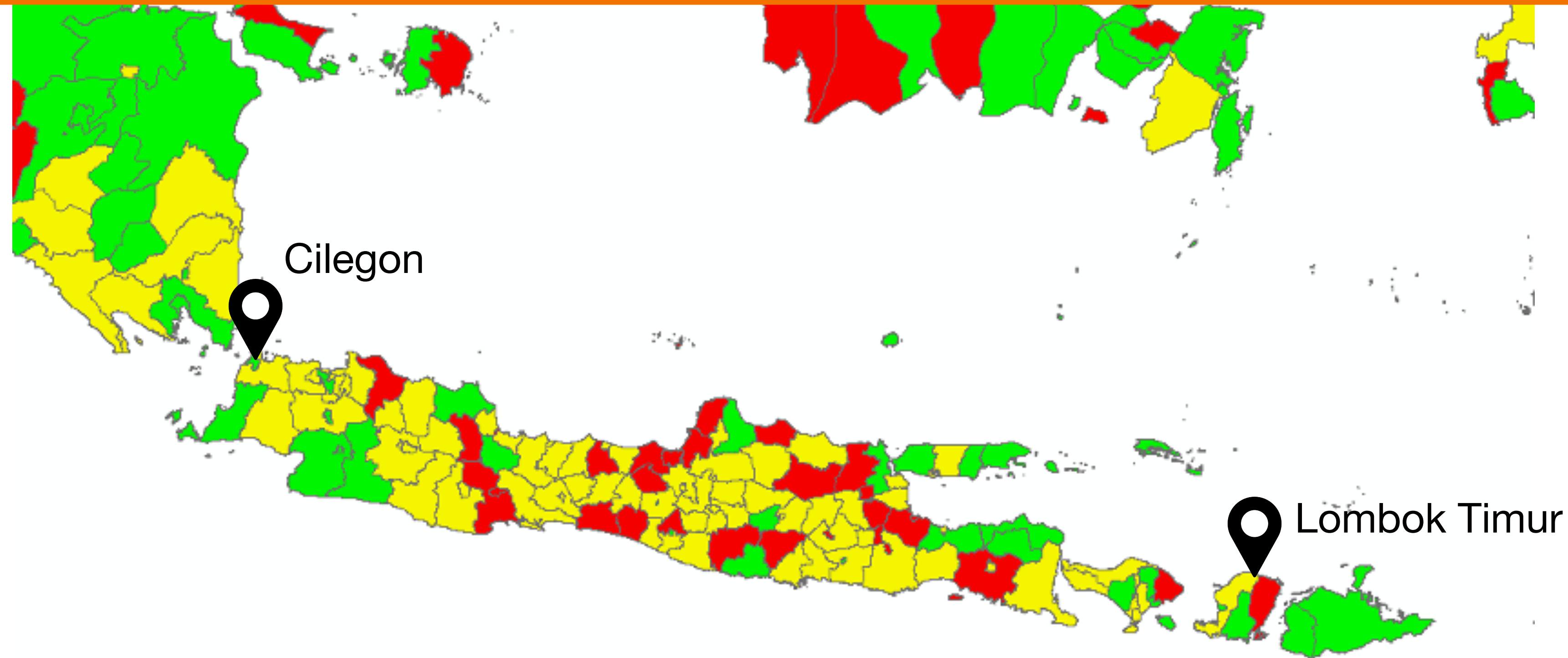
✳ The centroid of each cluster is 5.52, 5.58, and 5.66. Red is for the district with the lowest subjective happiness index, and green is for the highest index.

✳ The subjective happiness index can capture natural phenomena such as earthquakes. In Lombok Timur, Lombok citizens posted sadness, anxiety, and fear tweets during the streaming period due to the earthquake impact.

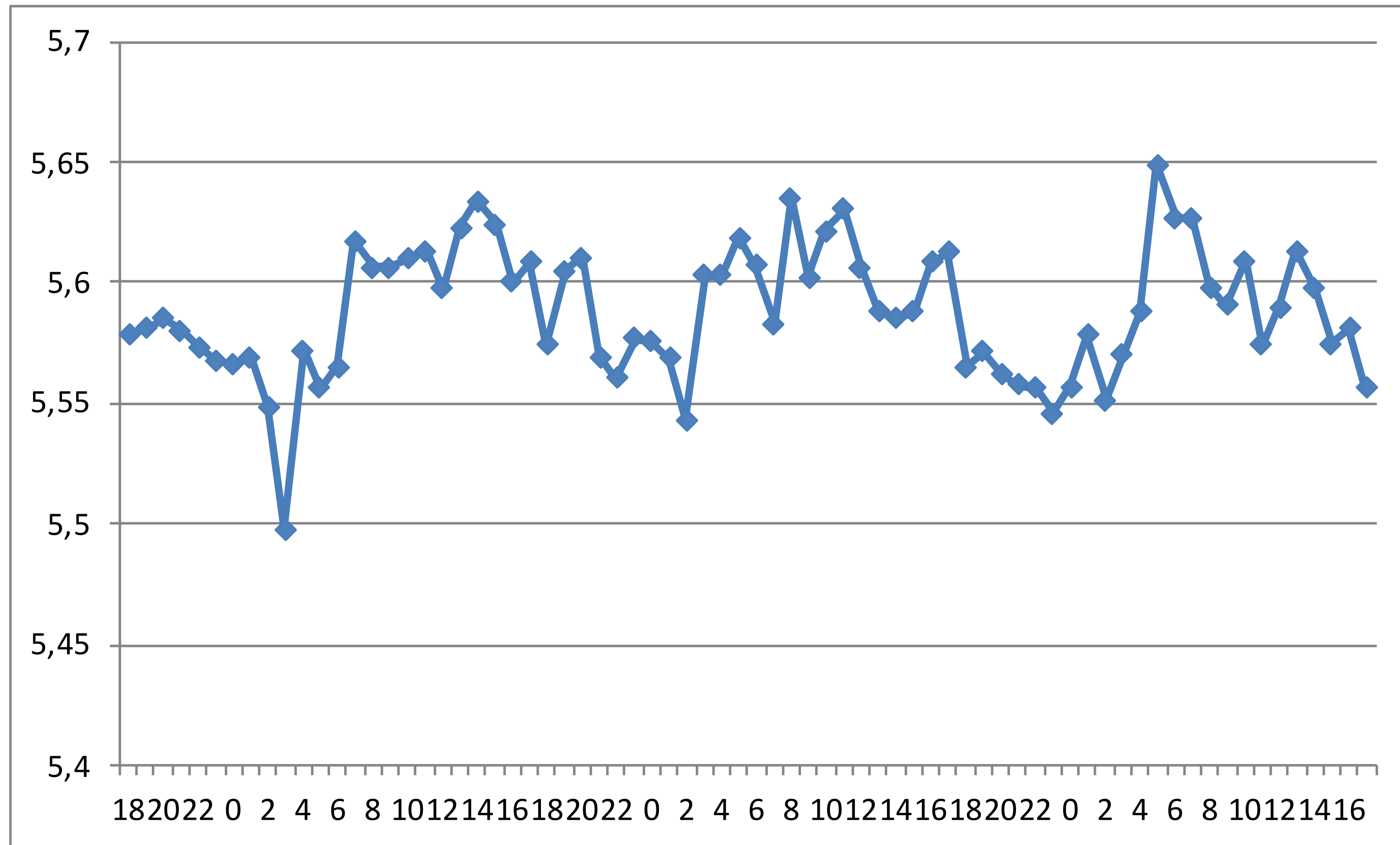# Number of Tweets per Hour

✦ The number of tweets posted each day shows a similar and consistent pattern day by day.

✦ The number of tweets reaches a pick at 12 noon and night (10 pm)

Trend of Subjective Happiness Index

❊ The subjective happiness index fluctuates per day but seems to show a similar pattern between days.

❊ Subjective happiness index decreases in the afternoon during going home from work and increases in the morning. It is higher at the weekend than on weekdays.

Conclusion

- Tweets reflect daily life and show emotion and expression of the public, whoever they are.

- Written in a formal Bahasa Indonesia and traditional and foreign languages, engagement with natural language processing technology is recommended.

- Given the sampling issue, this subjective happiness index still could be a signal or a complementary of a similar measurement. We can provide the subjective happiness index in more granular regions.

- The subjective happiness index tends to reflect a contrary pattern with the poverty rate.

- The average happiness scores match with Indonesian people' assessment through evaluation.

Thank you

Terimakasih