# Measuring well-being through social media: obstacles and opportunities

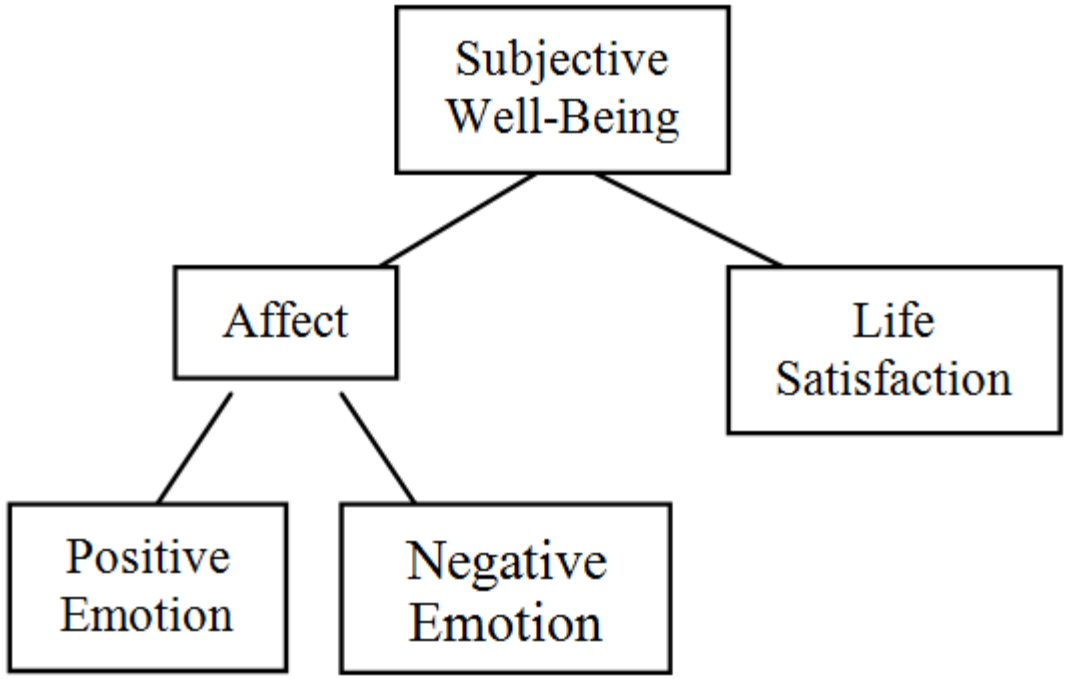
## Part I – GNH.today Index

Presented by Talita Greyling
(talitag@uj.ac.za)

THE EUROPEAN MASTER IN OFFICIAL STATISTICS (EMOS) – 13 March 2024

# Introduction


- Subjective well-being: definition and measures (SWB)
- Big Data and Social Media
- GNH.today Project – construction of real-time high-frequency time-series data
- Validation of GNH.today data
- Examples using GNH.today data to explain trends in happiness
- Future developments
- Application of the GNH.today data: “Vaccination, happiness and emotions: using a supervised machine learning approach”
- Conclusion




# Subjective well-being measures

Life satisfaction

- How satisfied are you with your life in general?

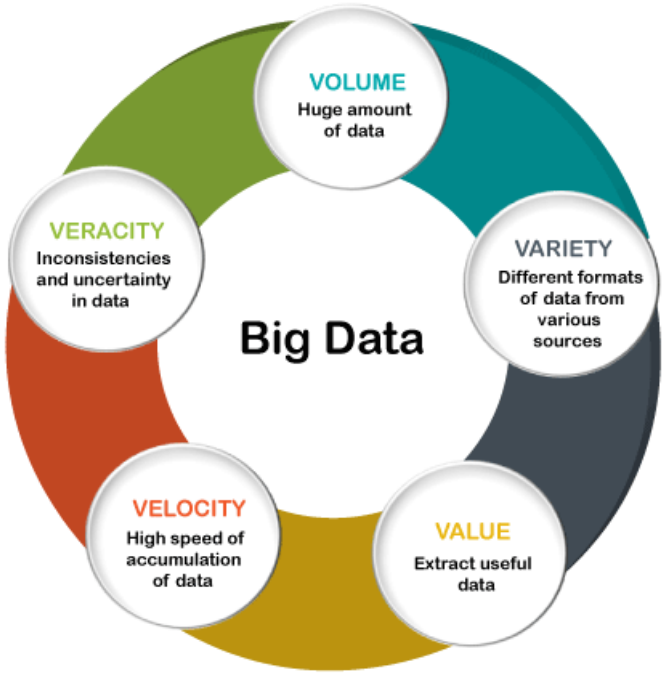
Positive Affect

- Enjoying life
- Loving others

Negative Affect

- Chronic worries
- Often angry

# What is Big Data?





**Data that are:**

- ***voluminous*** set of information
- collected at a high ***velocity***
- from ***various*** sources

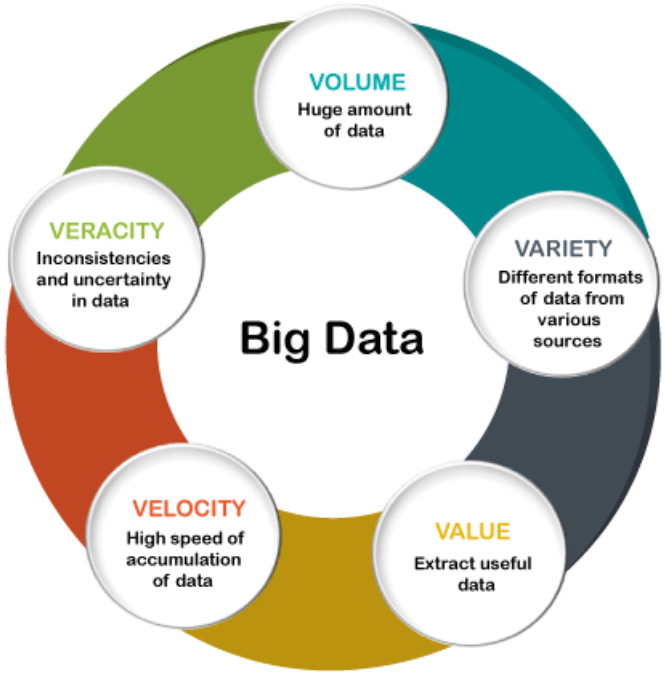
**Sources:**

- Health sector databases
- Financial institutions databases
- Google searches
- Social Media

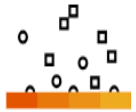
**Types:**

Structured, unstructured, and semi-structured datasets

# Social media data – Big Data




Structured data



Unstructured data

Semi-structured data

**Social Media Data** is generated by

- posts
- images
- videos
- comments

It gives opinions of people – ideal for subjective well-being measures


# Examples of Social media platforms - number of users

Facebook – Meta- 3,5 bil

Twitter (X) – Musk – 450 mil

Instagram – Meta – 2 bil

WhatsApp – 2 bil

TikTok – 1 bil

Reddit - 430 mil

Pinterest - 450 mil

Snapchat – 750 mil

YouTube - 2,5 bil

Google Trends Users ???

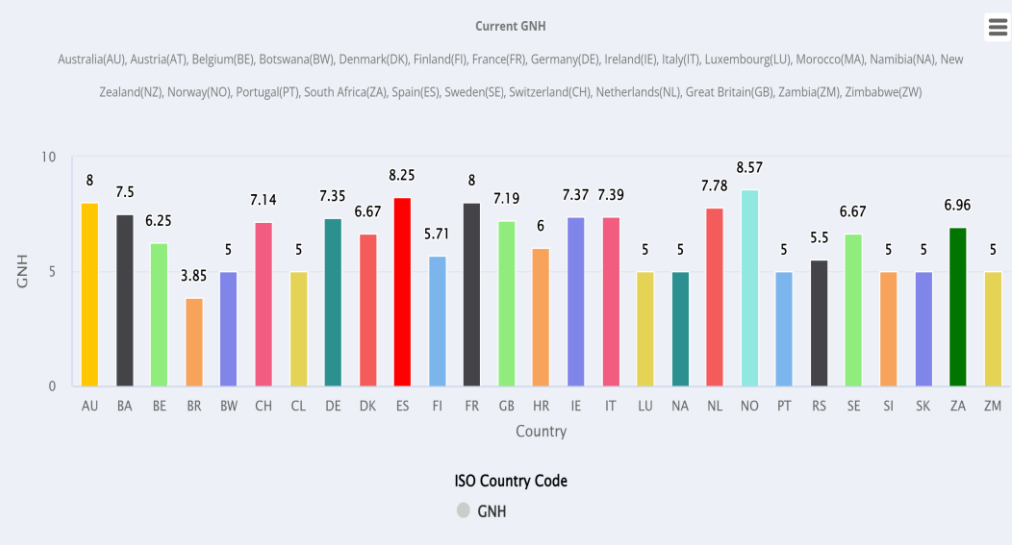

# Advantages of Big Data/ Social Media Data

- **Reduced costs** and efficient collection compared to costly survey collection and administration
- **Real-time** analysis and decision-making versus lagged information sharing (surveys)
- **Volume** – large datasets versus smaller survey datasets
- **Continuous updating** versus a snapshot in time
- **Unbiased and passive** data collection versus direct responses collected in surveys
- Attain **behavioural insights – social media listening** without the biases found in survey data
- Many more….

# Limitations of Big Data/ Social Media Data


- **Not a representative** sample; But can adjust the data to be more representative (see Iacus & Porro, 2021).
- To analyse Big Data, we need **sophisticated methodologies**.
- Big Data is **not always a substitute for surveys**, as surveys can provide context, nuance, and specific information that may not be captured in large-scale datasets.
- The **quality of the data** is crucial. Garbage in, garbage out.

# Gross National Happiness.today (GNH.today)




**What?**

High frequency near to real-time time-series data on happiness and emotions of populations

**Countries:**

Was ten countries:
Australia, New Zealand, South Africa, Belgium, Germany, Great Britain, France, Italy, the Netherlands, Spain.

**FIFA World Soccer Cup 2022:**

16 Countries – Argentina, Brazil, Morocco, Croatia, Uruguay, Portugal

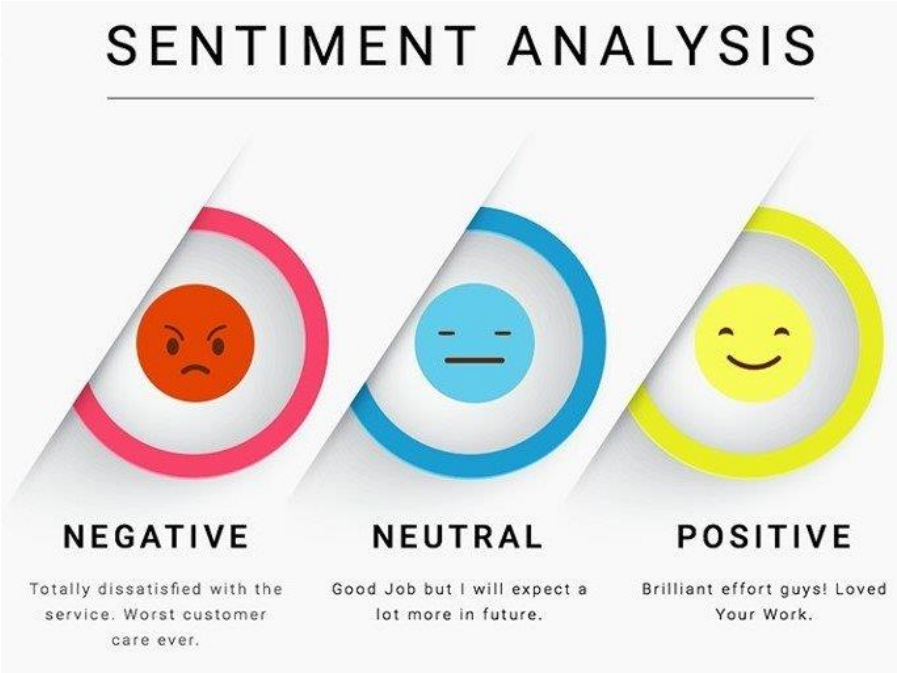
# GNH.today – the use of NLP – to determine sentiment  - the key of GNH.today

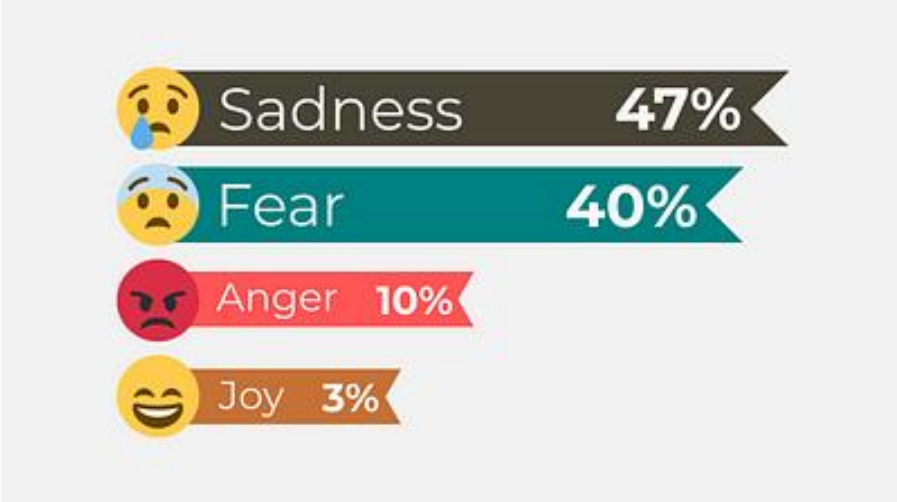
**What is a lexicon?**

Library of words – in NLP – a database that a computer program uses to understand the meaning and sentiment  of words.

**Why use lexicon - rule-based sentiment analysis (rather than a machine learning algorithm)**

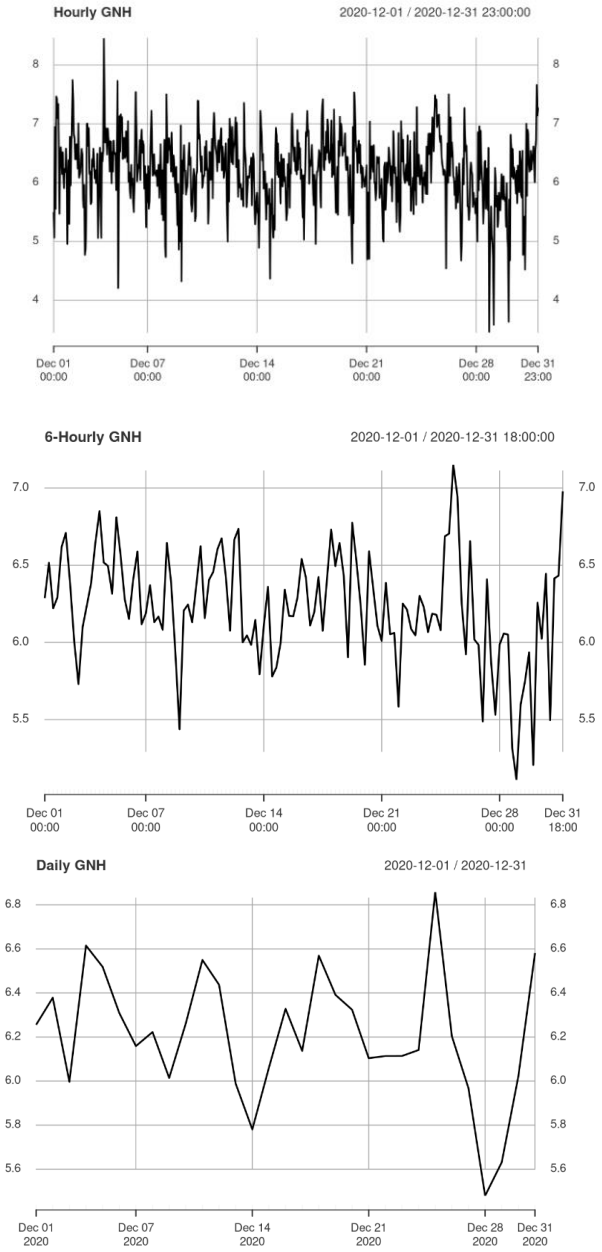
- straightforward
- less computationally expensive
- It doesn't require a large dataset for training,
- suitable for high-frequency analysis in real-time
- easier interpretation of how sentiment scores are derived,
- useful in understanding the rationale behind sentiment scores




# GNH.today - How?

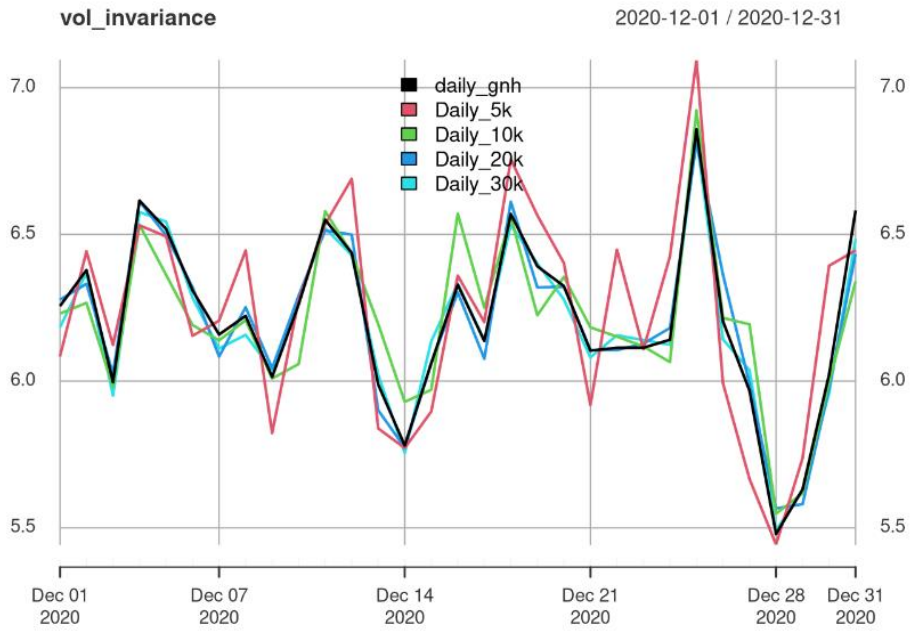
- Extract tweets in real-time – +/- 1 mil per day – geo-located (UTF-8)
- Determine languages – in ten countries – 64 languages. Translate – if needed
- Use lexicons to determine Sentiment/emotion analysis on the live stream of tweets
- Lexicons: Sentiment140, SYZHUET – bouquet, VADER, – further R&D –
- Average scores per hour/day/
- derive index using a balance equation
- Index varies between 0-10
- ***Accepted as official data by the New Zealand Statistical Services***




# GNH.today - Internal Validity

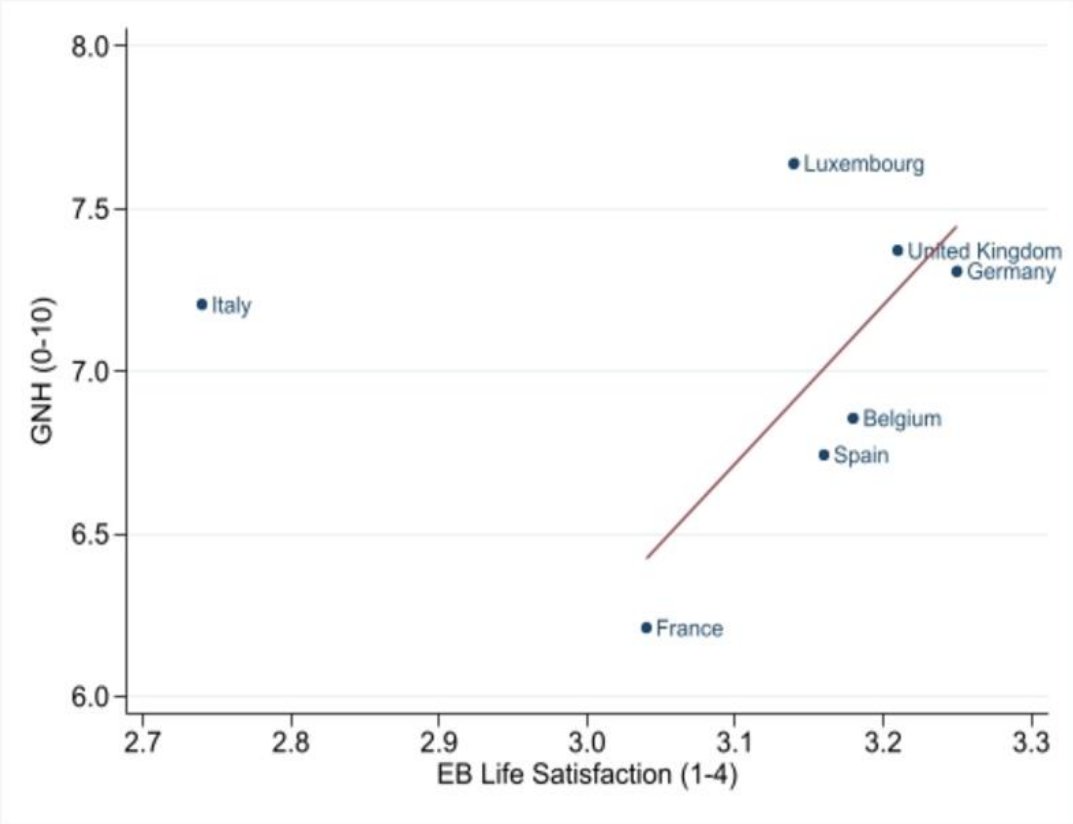
- **<u>Using timescale invariance</u>** to determine whether the sampling period significantly influences the GNH values.
- GNH.today is normally calculated per hour-
- what if we use averages per 6 hours or per day

# GNH.today – Internal Validity




- **<u>Using volume invariance</u>**
- To test if the volume of the sample directly influences the index, we sample 5k, 10k, 20k, and 30k tweets from the data extracted daily.

# GNH.today – Validity

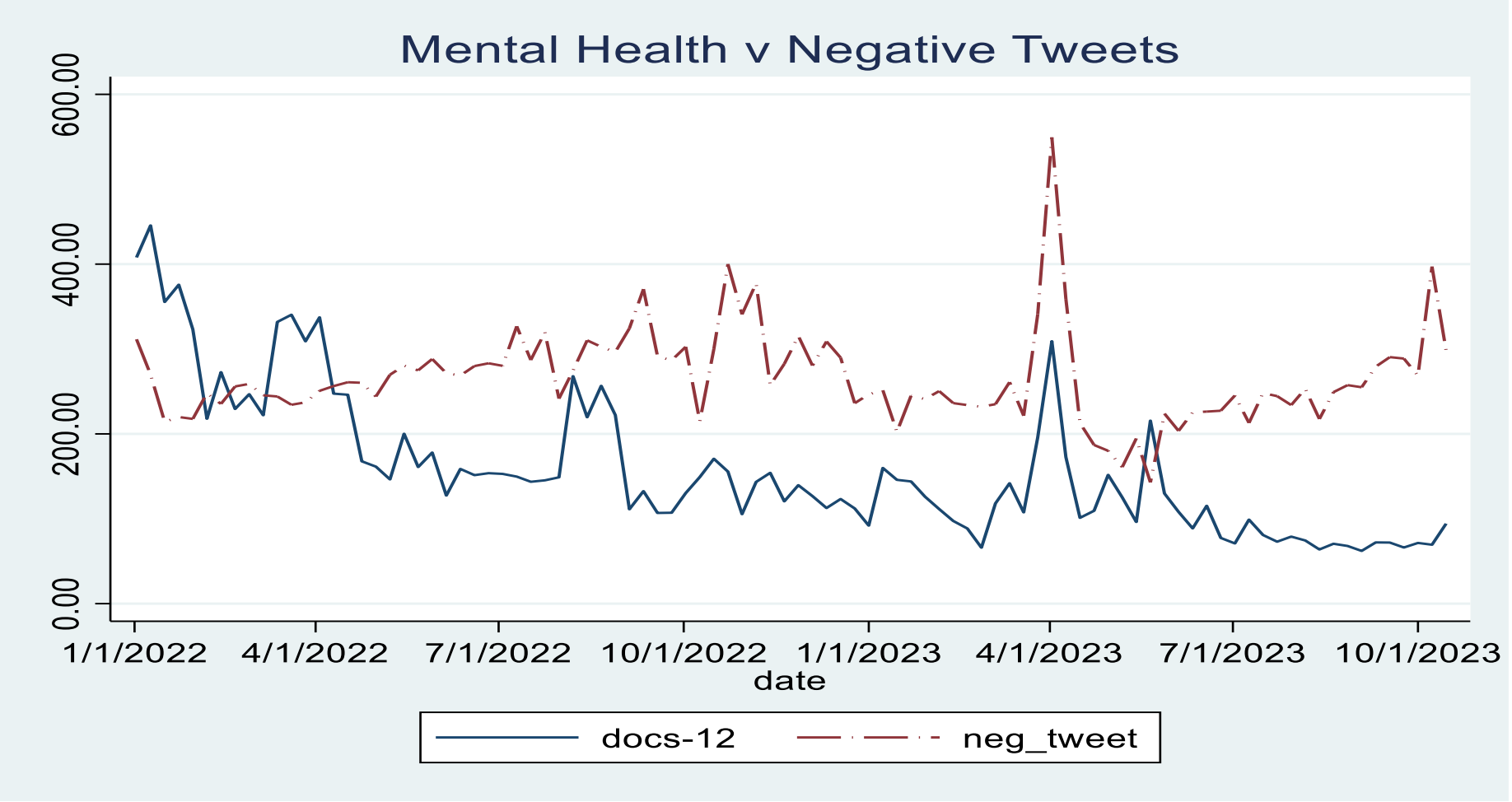



Mean **GNH correlates** positively with average life satisfaction of the **Eurobarometer** (Eurobarometer, Summer 2020).

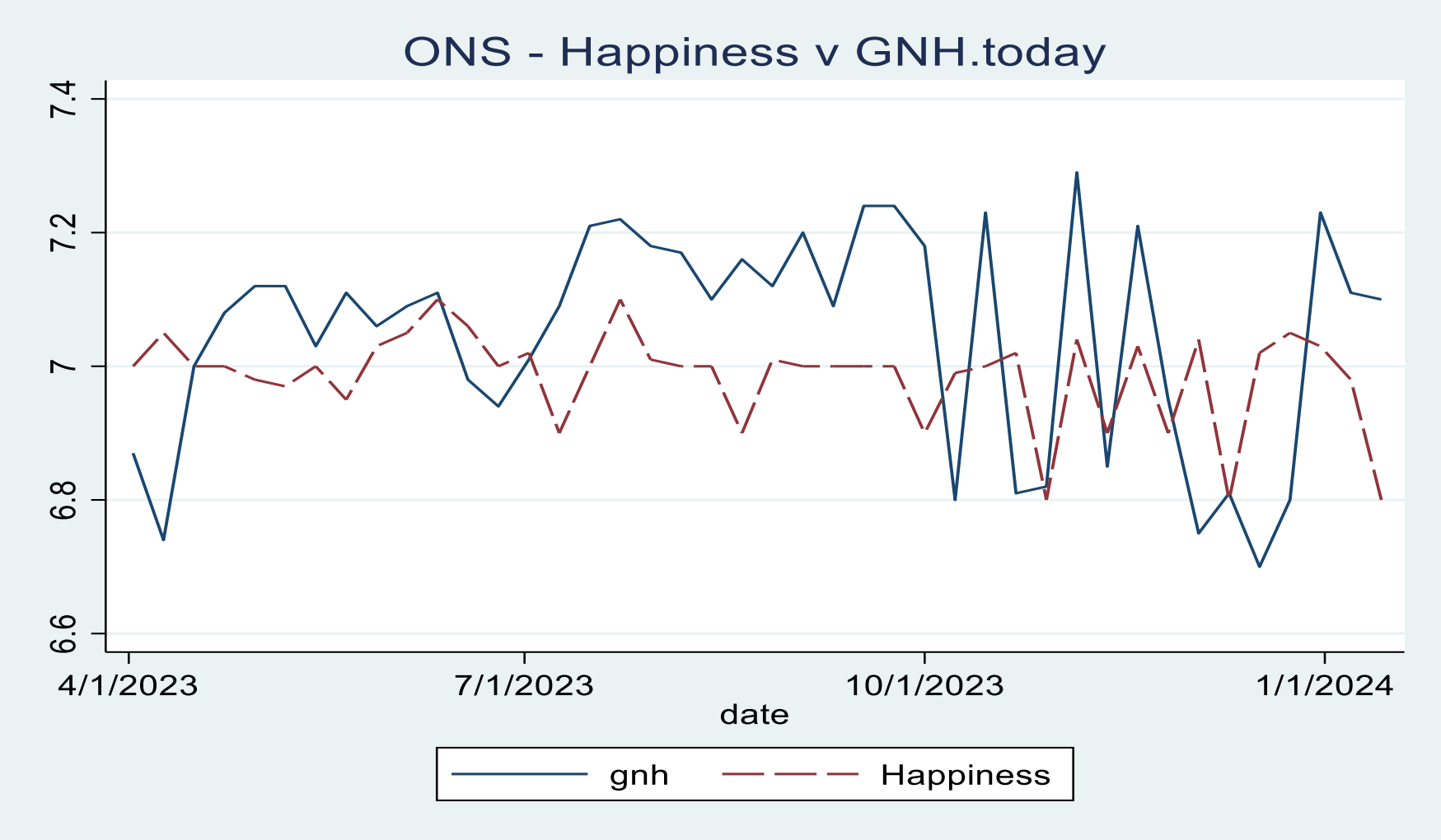
The GNH score is the average by country over the same period that the Eurobarometer was collected, from 9 July to 26 August 2020.

# GNH.today – Validity (Tweets v EARS (WHO) - Mental Health)

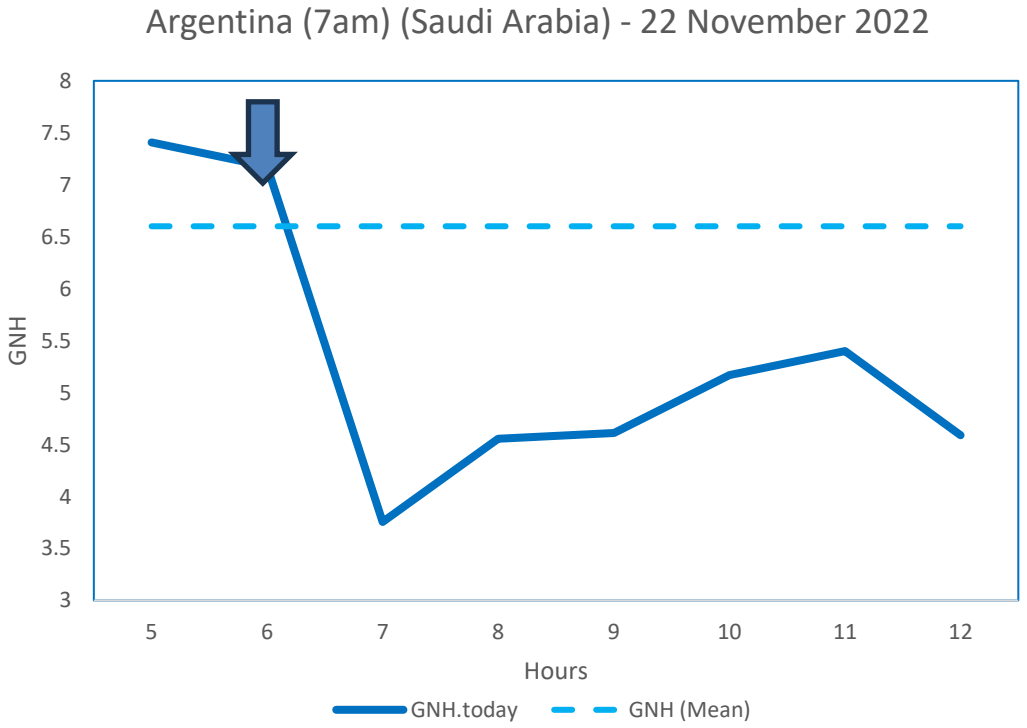
[EARS, Early AI-supported Response]



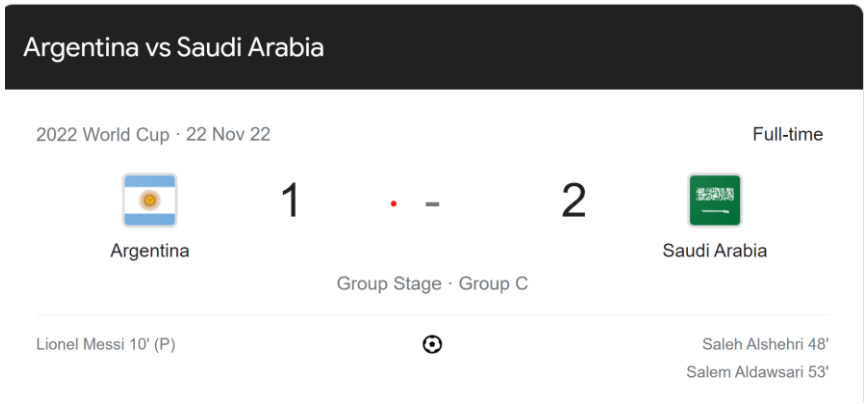
# GNH.today – Validity (GNH v Happiness - ONS)

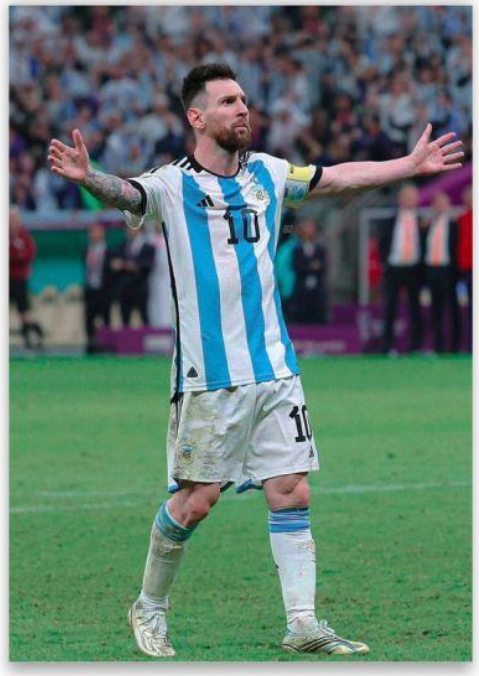


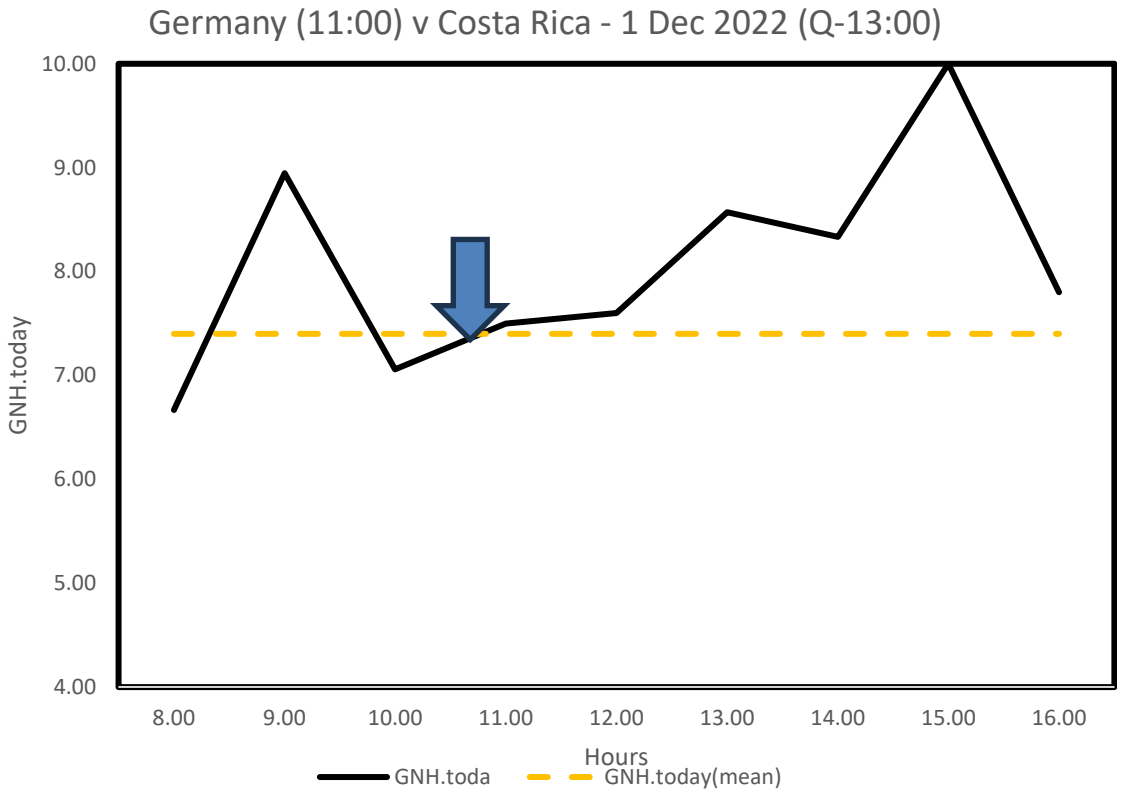
# GNH.today – validity- Examples

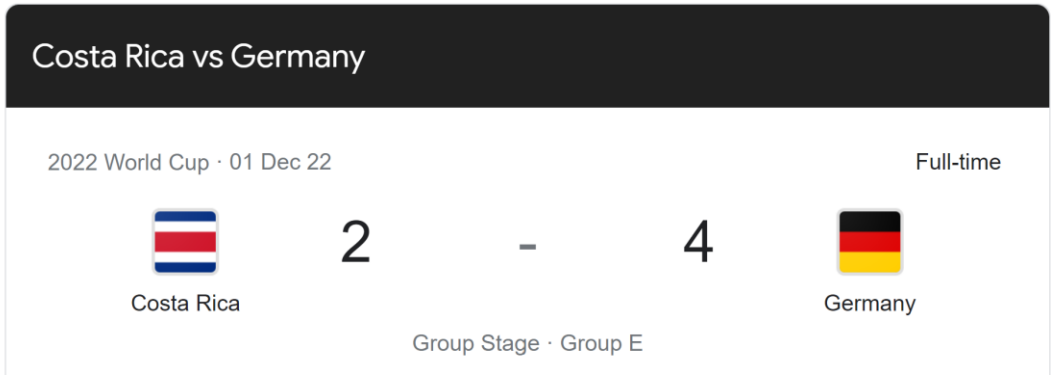



FIFA World Cup 2022 Argentina (6 am) – Saudi Arabia (Q -1pm)



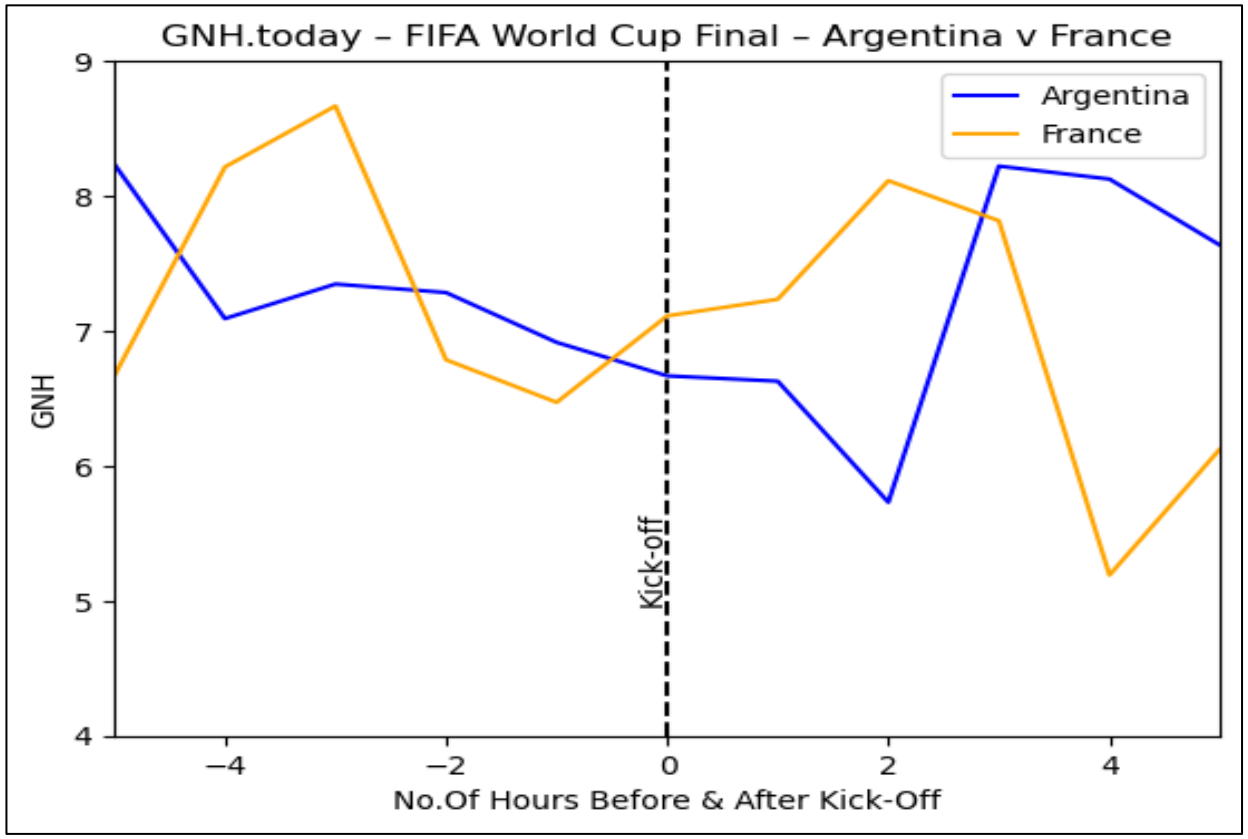



# GNH.today – validity - Examples

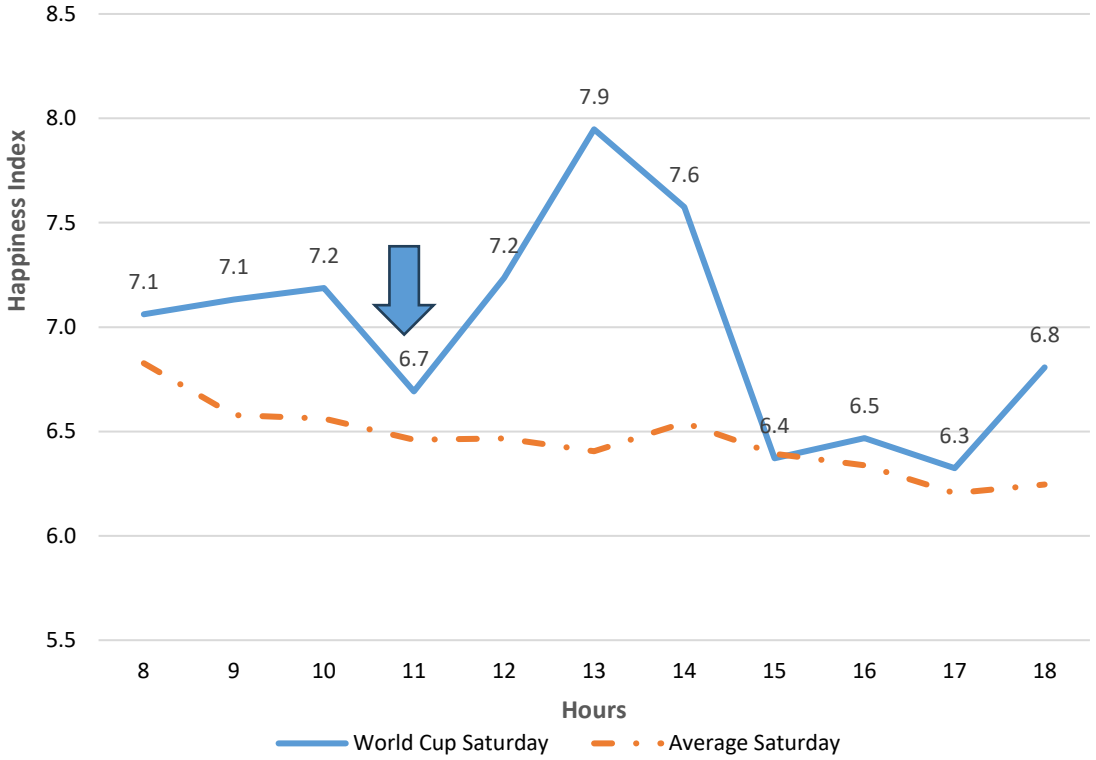
FIFA World Cup 2022 Germany (11 am) – Costa Rica

# GNH.today – Validity - Examples




FIFA World Cup Final – Argentina v France 2022

# GNH.today – Validity - Examples




World Cup – Rugby – South Africa win

# GNH.today – Future R&D


**Twitter:**

- Twitter (Elon Musk) closed the book on Twitter Research.
- This means the end of an era of research projects which contributed a wealth of knowledge on human behaviour.

**New projects:**

- Facebook
- Google Trends

# Application of the GNH.today data

## Vaccination, happiness and emotions: using a supervised machine learning approach. (Greyling, Greyling & Rossouw, 2023)

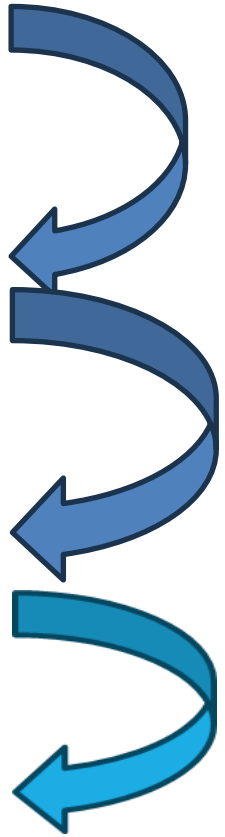
The COVID-19 pandemic is an example of an immense global failure to curb the spread of a pathogen and save lives.

770 million cases and almost 7 million deaths

**The problem:**

The failure could have been avoided if we reached herd immunity. Herd immunity is achieved from previous infection or through vaccination.
Vaccination is the method of choice (WHO).

To achieve herd immunity, early estimates were a  threshold of 70% of the population. However, because of mutation and infectiousness, estimates changed to 90%.



# Aim of the Study

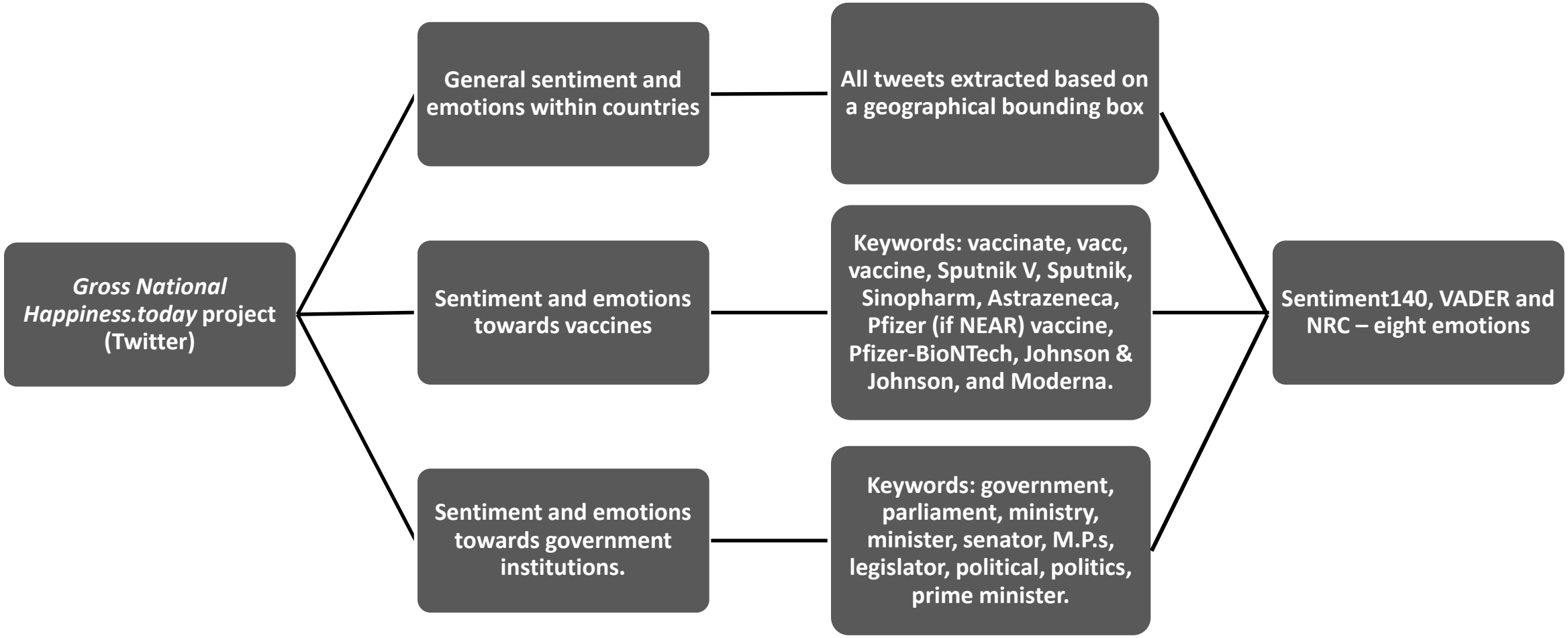
1. Retrospective evaluation of the COVID-19 pandemic – especially concerning vaccinations - to determine the most important factors to reach herd immunity at the 70% and 90% vaccination threshold seen as the “golden standard”.
2. Determine whether subjective well-being measures contribute to higher vaccine uptake. Why?
    - We know that decisions are driven by emotions – not the rational man theory

# Data - 1

1. Period under consideration – 1 December 2020 to 16 September 2022.
2. Ten countries – Northern and Southern hemispheres.
3. Four merged datasets – Google COVID-19 Open Data, 3 datasets from GNH.today

| |
|---|
| **Australia** |
| **Belgium** |
| **Germany** |
| **Spain** |
| **France** |
| **Great Britain** |
| **Italy** |
| **The Netherlands** |
| **New Zealand** |
| **South Africa** |

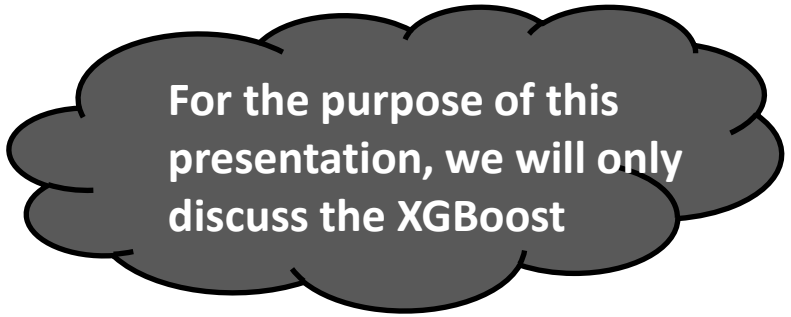
# Data - 2


*Gross National Happiness.today* project (Twitter)
General sentiment and emotions within countries
Sentiment and emotions towards vaccines
Sentiment and emotions towards government institutions.
All tweets extracted based on a geographical bounding box
Keywords: vaccinate, vacc, vaccine, Sputnik V, Sputnik, Sinopharm, Astrazeneca, Pfizer (if NEAR) vaccine, Pfizer-BioNTech, Johnson & Johnson, and Moderna.
Keywords: government, parliament, ministry, minister, senator, M.P.s, legislator, political, politics, prime minister.
Sentiment140, VADER and NRC – eight emotions

# Methodology - 1

Supervised machine learning in the form of:

1. XGBoost algorithm - tree-based ensemble method - implements gradient boosting.
2. More efficient
   - Computationally much lighter
   - Outperforms most supervised algorithms
3. Random Forest and Decision Tree as robustness tests.



# Methodology - 5

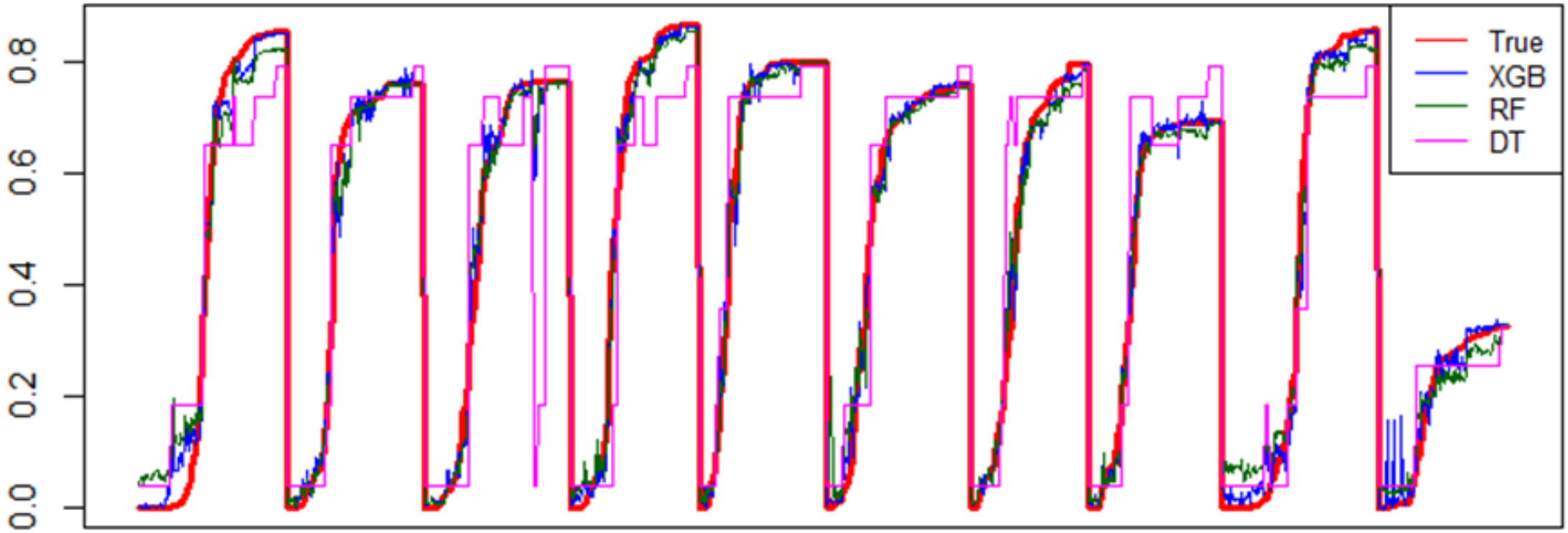
## Step 1: Training the model

- Total data is split into a training and testing dataset - 80:20 split.
- XGBoost model uses a maximum tree depth of seven. Run for 100 iterations with an early stop if the RMSE is not improved after five iterations. The model converges after 16 iterations.
- The Random Forest model. After 50 trees, it seemed as though the model converged, but upon further inspection of the model, the results continued to improve with minute increments with each additional iteration. We limit the number of iterations to 200.
- The Decision Tree model was easier to train since we have one tree with 8 nodes.

# Results – Model evaluation 1

| Model | MSE | MAE | RMSE |
|---|---|---|---|
| **XGBoost** | **0.001412552** | **0.022707714** | **0.0375839** |
| Random Forest | 0.001861686 | 0.029981258 | 0.043147264 |
| Decision tree | 0.01222601 | 0.07180425 | 0.11057130 |

# Results – Model evaluation 2




True value with all model predictions

# Results - interpretation of results based on XGBoost

| 90% vaccine threshold | 70% vaccine threshold |
|---|---|
| **Vaccination policy** | **Vaccination policy** |
| **International travel controls** | Population aged between 10-19 |
| **Percentage of population in rural areas** | **International travel controls** |
| **Happiness** | **Percentage of population in rural areas** |
| **Average temperature** | **Average temperature** |
| Population density | Workplace closing |
| Human Development Index | Restrictions on gatherings |
| Facial coverings | Life expectancy |
| Workplace closing | **Happiness** |
| Restrictions on gatherings | Pollution mortality rate |

# Conclusions – of Vaccines and happiness paper

1. Overlap in factors 70% threshold and 90% threshold:
   - Vaccination policy implemented,
   - International travel controls,
   - The percentage of the population in rural areas
   - Average temperature.
2. Happiness plays a major role in achieving a 90% vaccination threshold:
   - To reach 70% threshold can be achieved with policy measures (objective policy interventions). (Happiness 9$^{th}$)
   - To reach 90% we need to consider subjective measures – the mood of a nation. (Happiness 4$^{th}$)
3. If governments want higher levels of compliance and vaccine uptake, subjective well-being measures such as mood and emotions must be prioritised.

# A few published papers using GNH.today data


Greyling T, Rossouw S . 2024. Reactions to macro-level shocks and re-examination of adaptation theory using Big Data. PLoS ONE 19(1): e0295896. https://doi.org/10.1371/journal.pone.0295896 **Greyling, T.** & Rossouw, S . 2024.

Sarracino, F., **Greyling, T**., O'Connor, K., Peroni, C. & Rossouw, S. 2023. A Year of Pandemic: Levels, Changes and Validity of Well-being Data from Twitter. Evidence From Ten Countries, *PLOS ONE*, *18*(2), e0275028. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0275028

**Greyling, T.,** Rossouw, S & Steyn, D. H. W. 2022. The Prediction of Intra-Day Stock Market Movements in Developed and Emerging Markets using Sentiment and Emotions from Twitter**.** *Finance India*. 24(3): 907-939.

Adhikari, T., **Greyling, T.** & Rossouw, S. 2022. The ugly truth about social welfare payments and households' subjective well-being. *South African Journal of Economic and Management Sciences, 25(1).*

**Greyling, T.** & Rossouw, S. 2022. Positive attitudes towards COVID-19 vaccines: A cross-country analysis. PLoS ONE 17(3): e0264994. https://doi.org/10.1371/journal.pone.0264994

**Greyling, T**., Rossouw, S. & Adhikari, T. 2021. The good, the bad and the ugly of lockdowns during Covid-19. PLoS ONE 16(1): e0245546. https://doi.org/10.1371/journal.pone.0245546

Rossouw, S., **Greyling, T.** & Adhikari, T. 2021. Happiness lost: Was the decision to implement lockdown the correct

# Conclusion

- Using social media data – we can capture the happiness and emotions of people in real-time
- Statistical offices should experiment with big data and implement data derived from Big Data and Social Media – it gives access to real-time data and can be used alongside survey data.

# Questions?