MINING SOCIAL MEDIA DATA: INFORMATION DIFFUSION AND SOCIAL ENGAGEMENT IN THE GREATER REGION

Ninghan Chen
UL-FSTM/DCS
May 20, 2021

AIM AND SCOPE

Since its outbreak, COVID-19 has become an unprecedented global health crisis. Pandemic information diffusion over online social media can strongly influence people's behaviour and significantly deteriorating the efficacy of measures taken by governments.

The goals of my research are as follows:

- Analysing user engagement, communication topics on Twitter during COVID-19.
- Developing a machine learning model to understand and predict COVID-19 information cascade over social networks.
- Figuring out what role do multilingual users play in diffusing COVID-19 related information on social media and how they are influenced in subjective well-being by social media information.

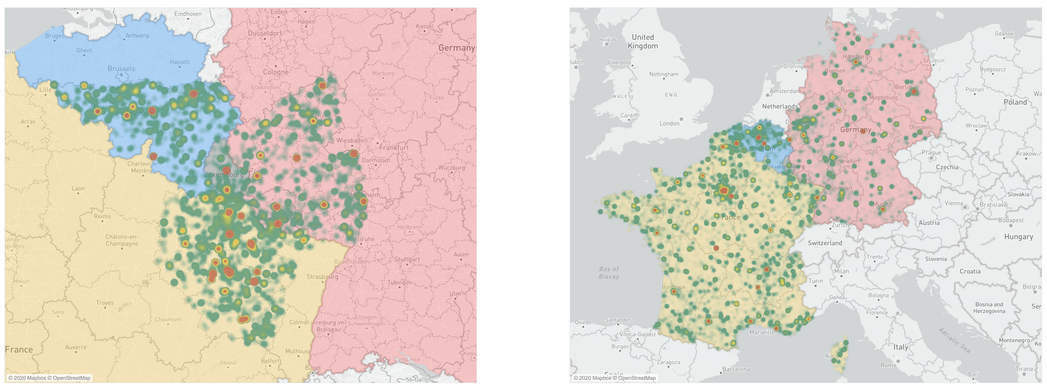
METHODOLOGY

The project took interdisciplinary approaches to social media research by combining data mining techniques and computational social science methodologies in The Greater Region(GR).

- Natural Language Processing methods to learn the latent information from text content.
- Deep Learning methods to learn node representations with neighbourhood aggregation from social network.

STUDY BACKGROUND

FIGURE 1



(a) GR (b) [The related countries

User location heatmap of GR and the related countries.

The Greater Region has the highest number of cross-border commuters in Europe, with approximately 250, 000 commuters per day. This makes GR a representative region with high mobility and multilingualism [1].

DATA SET

TABLE 1

Region/Country	tweet volume	User volume
Global	51,966,639	15,551,266
GR	35,329	7,894
Luxembourg	7,512	1,545
Belgium	119,467	31,446
France	1,050,312	288,009
Germany	430,688	87,796

Summary of our COVID-19 Twitter dataset.

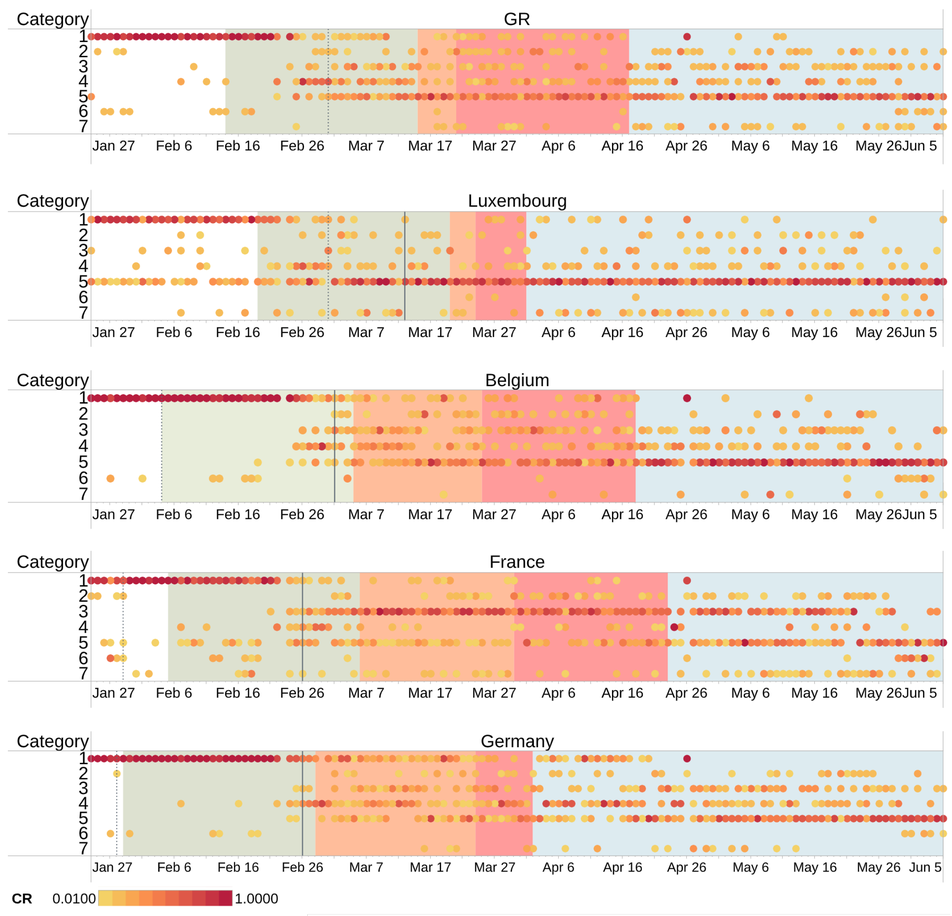
TABLE 2

#node	5,808,938
#edge	12,511,698
Average degree	2.15

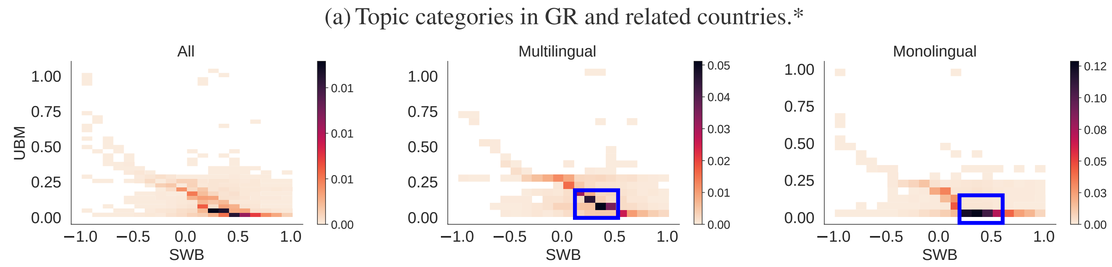
Statistics of the GR-ego social network.

EXPERIMENTAL RESULTS

FIGURE 2



(a) Topic categories in GR and related countries.*



(b) Distribution of Subjective well-being and bridging values on different user groups.

*1: Wuhan& China 2: Measures 3: Local news 4: International news 5: Policy and daily life 6: Racism 7: Other

TABLE 3

Evaluation Metric	MRSE	MAPE	WroPerc
Feature-based	0.3611	0.4018	41.31%
DeepCas	0.2837	0.3959	37.71%
CoupledGNN	0.2678	0.3861	35.19%
SE-CGNN	**0.2514**	**0.3803**	**33.65%**

Cascade prediction results.

1. User engagement and communication topics.
   - Tweets volume and COVID-19 cases are correlated, but this correlation only exists during the early period of the pandemic.
   - Sporadic cases may not attract enough public attention, and the public's attention was still focused on China-related news until a complete outbreak.
   - The public in Luxembourg and GR started to have discussions about measures 1-2 days before the first case appeared.
2. Information cascade predict model.
   - We build a model (SE-CGNN) to capture spillover effects and improve the accuracy of cascade prediction results of COVID-19 information on Twitter.
   - Information about unemployment and school closures have spillover effects on anti-virus measure information.
3. The burden of being a bridge: understanding the role of multilingual users.
   - Multilingual users have been playing an important bridging role in diffusing COVID-19 related information.
   - There is a negative correlation between information relaying behaviour and subjective well-being.

ACKNOWLEDGEMENT

The Doctoral Training Unit **Data-driven computational modelling and applications** (DRIVEN) is funded by the Luxembourg National Research Fund under the PRIDE programme (PRIDE17/12252781). https://driven.uni.lu

REFERENCES

[1] *The Greater Region at a Glance*.
http://www.granderegion.net/en/The-Greater-Region-at-a-Glance.