netherlands eScience center

PuReGoMe: Social Media Analysis of the Pandemic

Erik Tjong Kim Sang
Netherlands eScience Center
11 February 2021

- Cooperation with the University of Utrecht (Wang, Schraagen & Dastani)
- eScience project TwiNL (2012-2013) collected 4 billion Dutch tweets

Challenges

- No demographic information on Twitter users (TKS & Bos, 2011)
- Too few data to answer interesting questions
- Misinformation (e.g. actions of Cambridge Analytica and IRA in 2016 elections)

Request from funder

- Cooperate with similar project of Radboud/Leiden/RIVM (Van den Bosch/Verberne)
- The university involvement finished in July 2020

Project interests

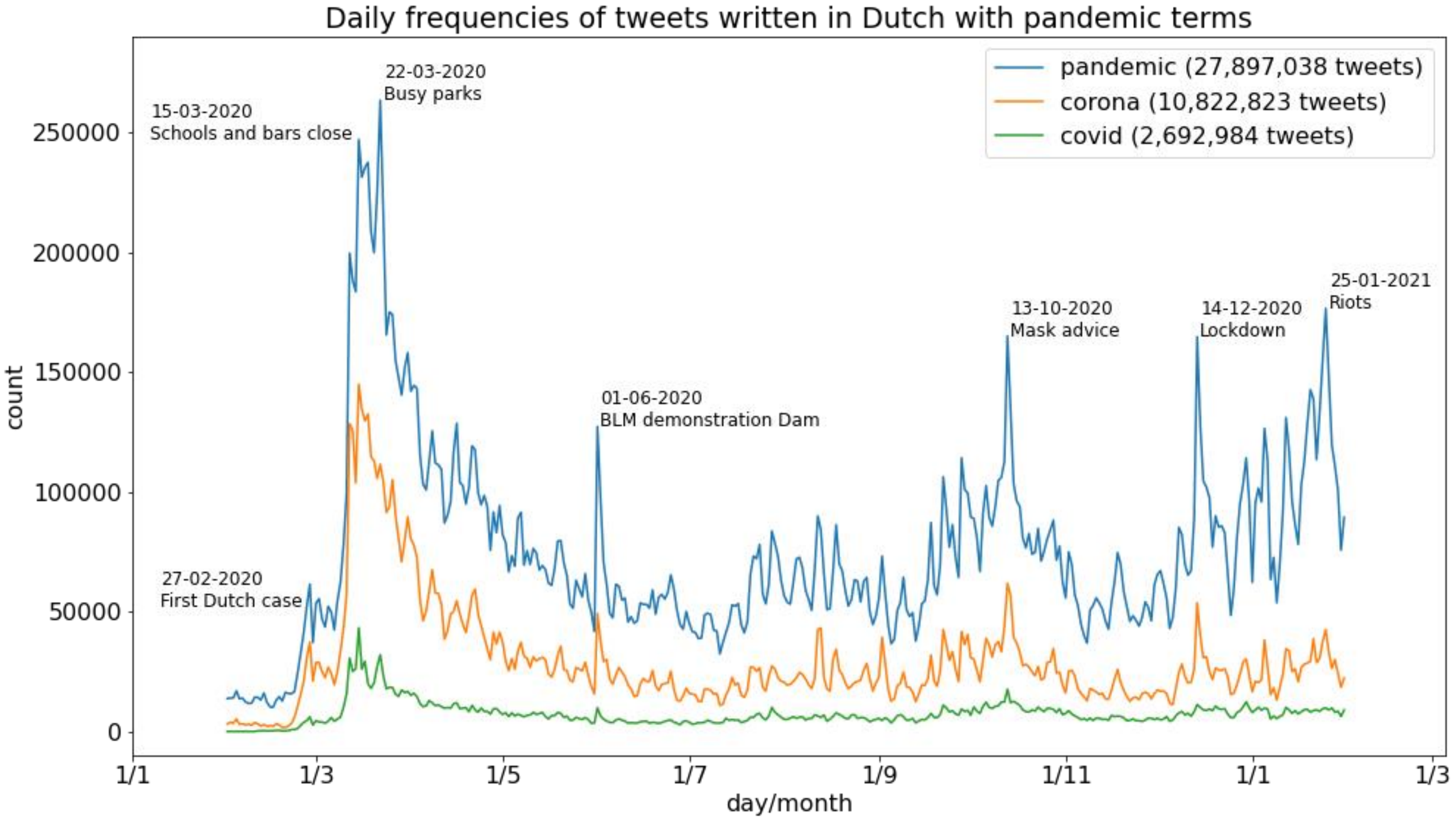
- Topics of discussion
- Frequencies of topics
- Sentiment of discussions
- Stances on pandemic measures
- Geographic variation of stances

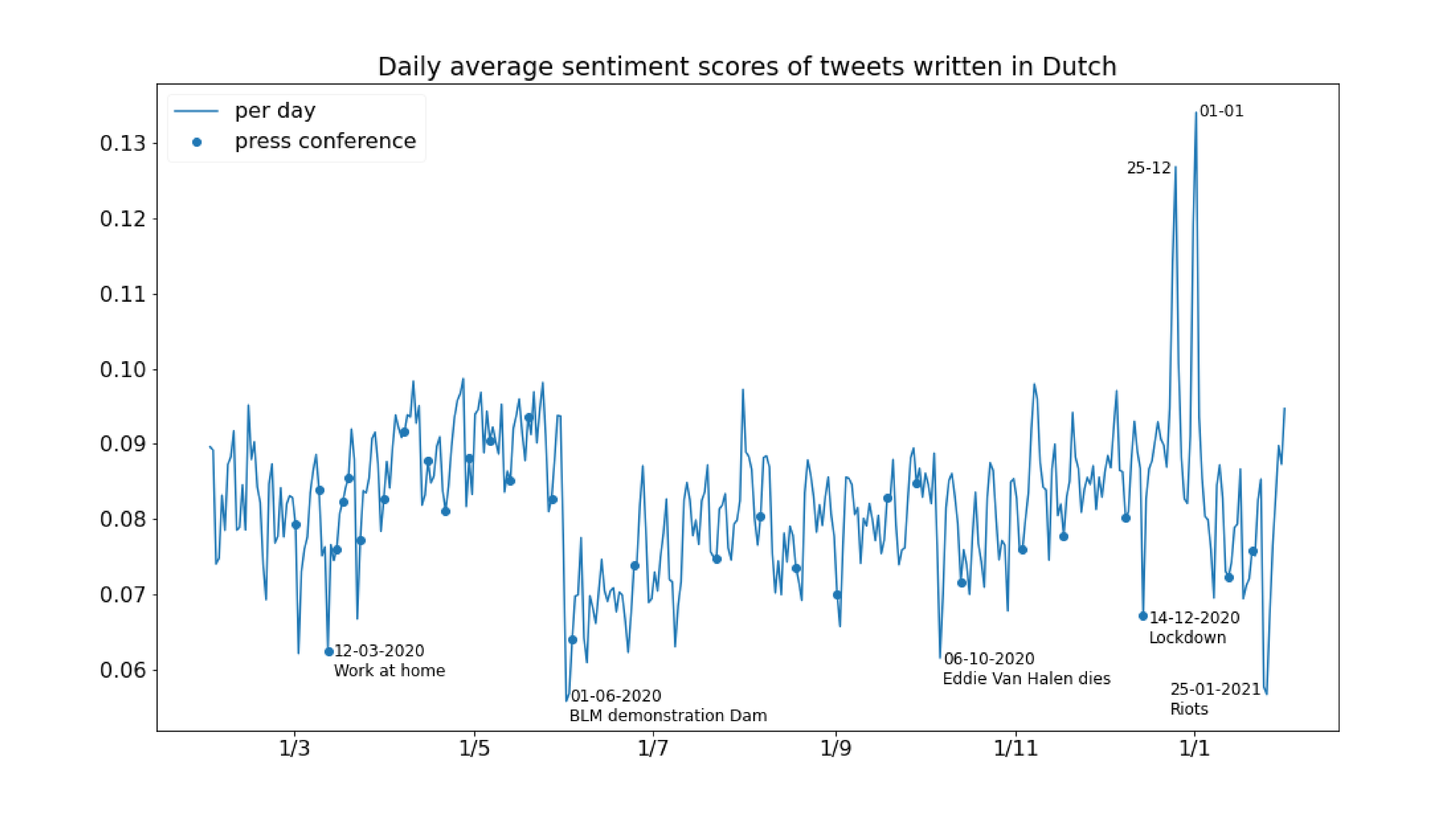
Scientific techniques and technologies used: Latent Dirichlet Allocation, *t*-test, keyword search, sentiment lexicons (Pattern), machine learning (FastText)

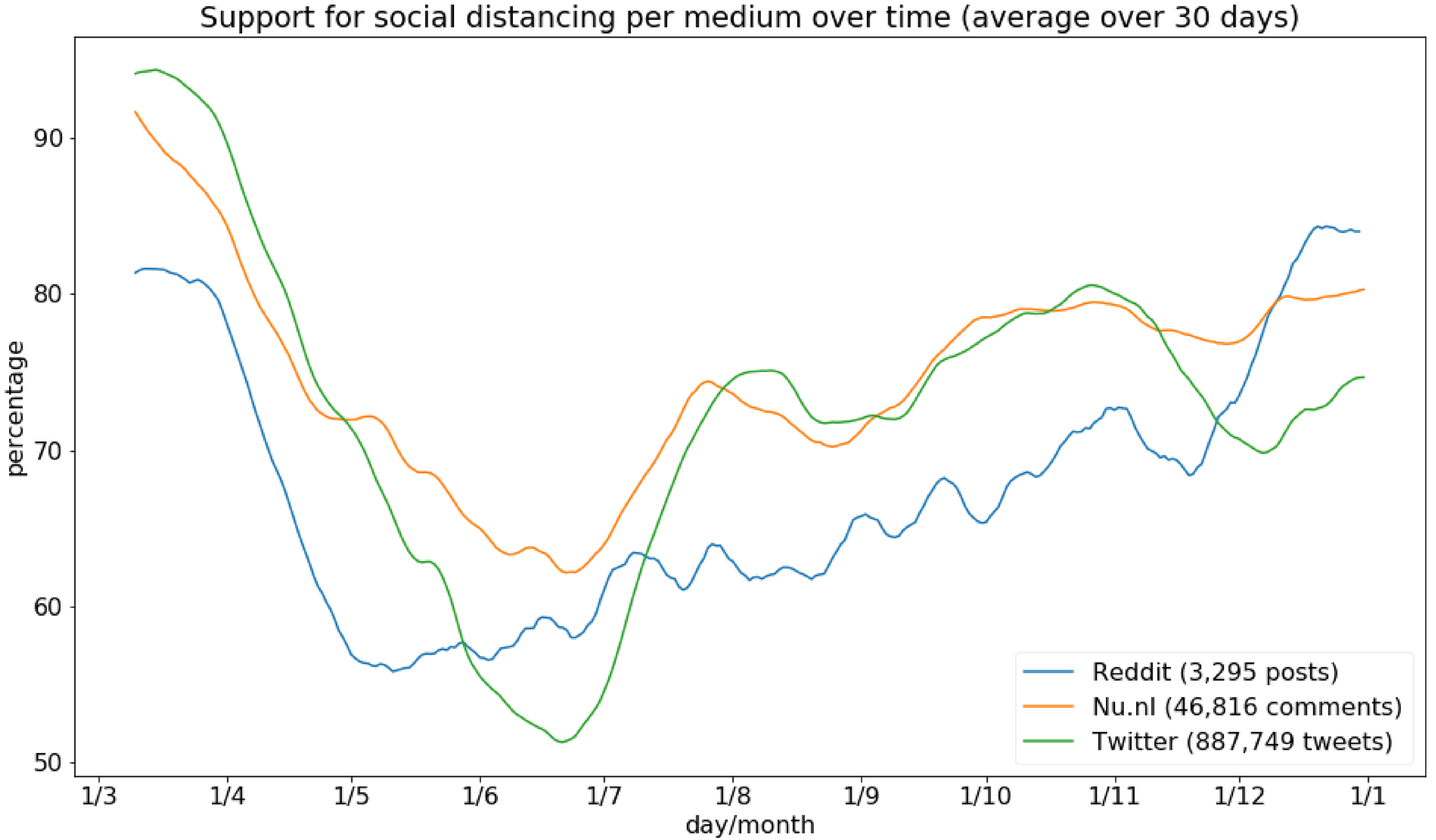
Relevant words for the pandemic topic beside *corona* and *covid*: they change all the time

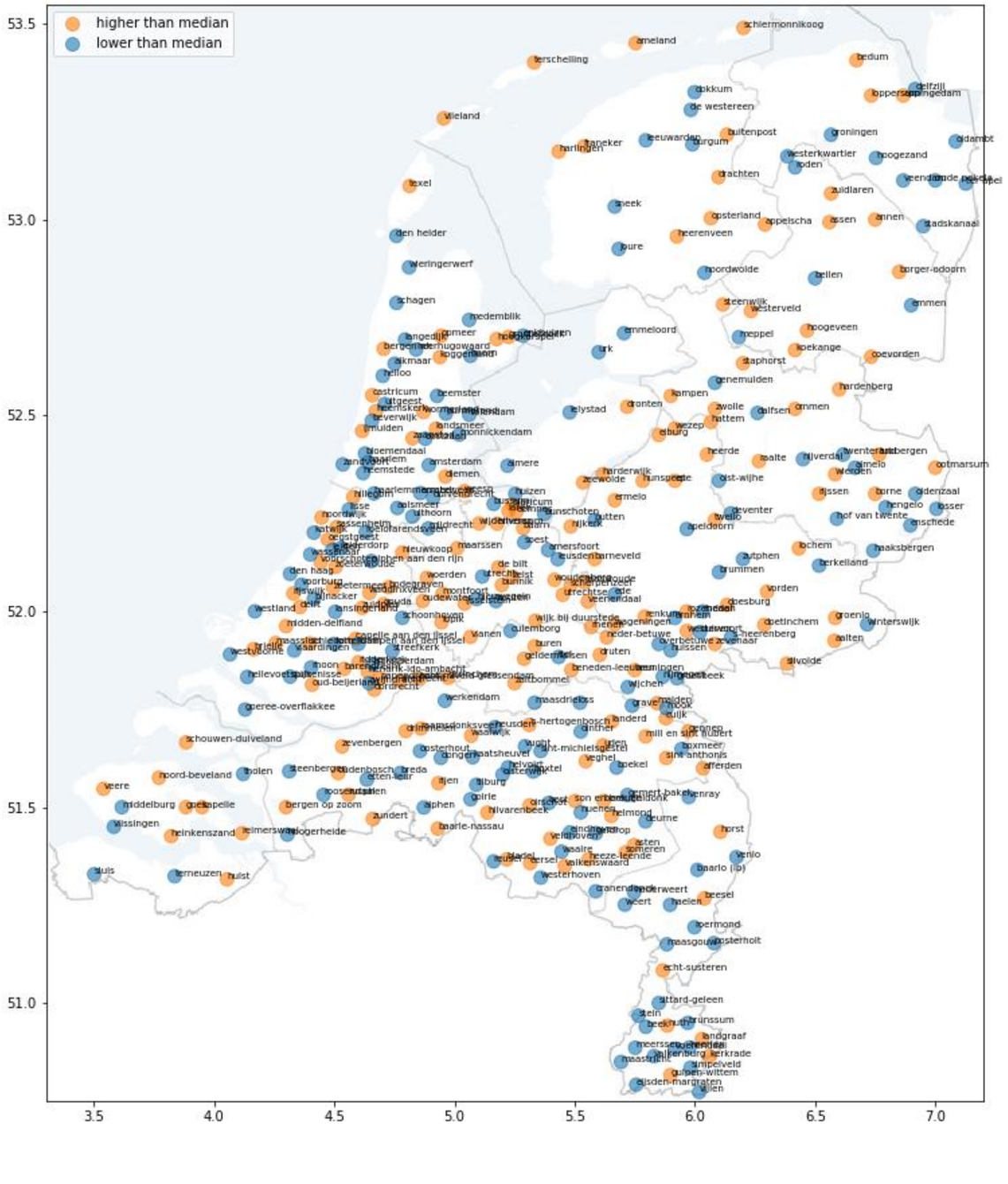
January-December 2020					January 2021
1. virus	16. ic	31. Italië	46. economie	61. positieve	1. maatregelen
2. crisis	17. overleden	32. klachten	47. GGD	62. maanden	2. virus
3. maatregelen	18. zorg	33. afstand	48. mondkapjes	63. overheid	3. debat
4. debat	19. positief	34. beleid	49. alleen samen	64. maart	4. crisis
5. aantal	20. vaccin	35. opgenomen	50. getroffen	65. wet	5. vaccin
6. patiënten	21. tijd	36. totaal	51. quarantaine	66. minister	6. **avondklok**
7. doden	22. test	37. artsen	52. huisarts	67. tweede	7. lockdown
8. Nederland	23. lockdown	38. uitbraak	53. controle	68. WHO	8. aantal
9. besmettingen	24. regels	39. ziek	54. nieuwe	69. miljoen	9. Britse
10. testen	25. aanpak	40. scholen	55. 1,5	70. strijd	10. mensen
11. RIVM	26. Rutte	41. ouderen	56. houden	71. sterven	11. **variant**
12. griep	27. cijfers	42. Hugo de Jonge	57. informatie	72. gevallen	12. patiënten
13. kabinet	28. China	43. persconferentie	58. impact	73. voorkomen	13. ic
14. ziekenhuis	29. gevolgen	44. pandemie	59. app	74. horeca	14. beleid
15. verspreiding	30. weken	45. blijf thuis	60. FvD	75. risico	15. griep

Green words: candidates for pandemic topic (others: symptomen, intensive, golf, verpleeghuizen, **Pfizer**)


Daily frequencies of tweets written in Dutch with pandemic terms
pandemic (27,897,038 tweets)
corona (10,822,823 tweets)
covid (2,692,984 tweets)
15-03-2020
Schools and bars close
22-03-2020
Busy parks
01-06-2020
BLM demonstration Dam
13-10-2020
Mask advice
14-12-2020
Lockdown
25-01-2021
Riots
27-02-2020
First Dutch case
count
0
50000
100000
150000
200000
250000
1/1
1/3
1/5
1/7
1/9
1/11
1/1
1/3
day/month


Daily average sentiment scores of tweets written in Dutch
per day
press conference
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
01-01
25-12
12-03-2020
Work at home
01-06-2020
BLM demonstration Dam
06-10-2020
Eddie Van Halen dies
14-12-2020
Lockdown
25-01-2021
Riots
1/3
1/5
1/7
1/9
1/11
1/1

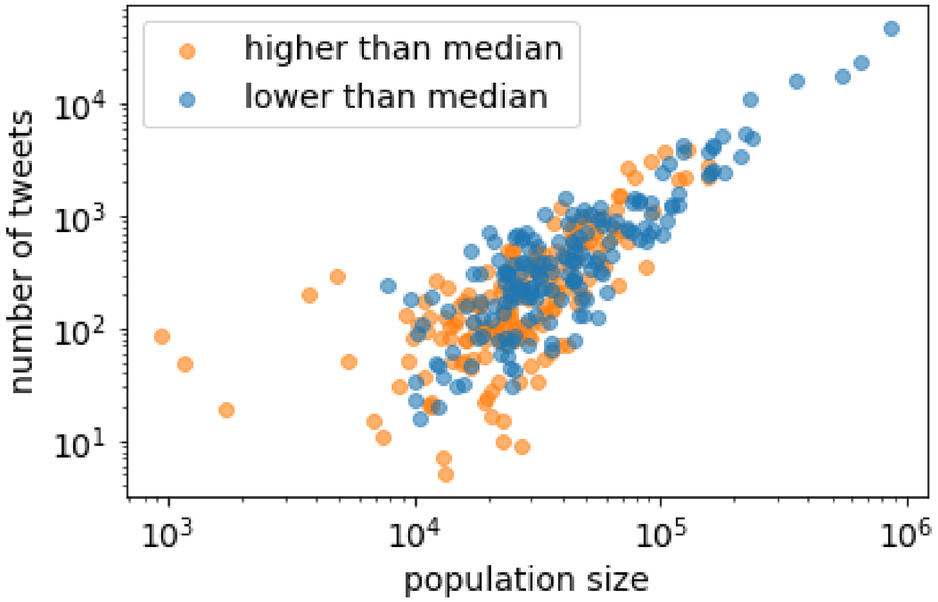

Support for social distancing per medium over time (average over 30 days)
percentage
90
80
70
60
50
1/3
1/4
1/5
1/6
1/7
1/8
1/9
1/10
1/11
1/12
1/1
day/month
Reddit (3,295 posts)
Nu.nl (46,816 comments)
Twitter (887,749 tweets)



Support for social distancing

Left map: no obvious regional differences

Bottom graph: possible effect of population size



Concluding remarks

- PuReGoMe has produced an interesting pandemic dataset (28M Dutch tweets)
- Despite concerns of data quality, we found some useful patterns in the data
- A constraint for the data analysis is the limited availability of data annotators
- We also ran out of project hours back in October

The goal of the project was to extract information from social media to support the decisions of the RIVM. However, RIVM prefers using questionnaires for retrieving this information. They used that data to create this website:

https://www.rivm.nl/gedragsonderzoek/maatregelen-welbevinden