Regional sentiments in Covid tweets in the Netherlands before and during peak infections

Tykhonov, Vyacheslav and van Leeuwen, Bas

DANS, IISH

December 2021

Regional sentiments in Covid tweets in the Netherlands before and during peak infections

Vyacheslav Tykhonov (DANS) and Bas van Leeuwen (International Institute of Social History)

Abstract

We use numbers of Covid-related tweets over Dutch regions in month 1-9 of 2020. Peaks in tweets precede the peaks of infections by about one month (the exception is June when the corona emergency law ("Spoedwet") was introduced, which drew a lot of online comments. The reason for this time lag is that more positive sentiments, which resulted in fewer tweets, occurred during peak infections. Just before, more negative sentiments dominated causing more tweets. This positivity in tweets during peak infections has, no doubt, various reasons. Yet, one reason may be in how people value society: the higher the number of infections, the more positive the sentiments related to crucial occupations became, which resulted in fewer tweets. This relation does not hold for non-crucial occupations.

1. Introduction

One of the biggest mysteries of alpha sciences such as economics and history concerns the sentiments of people. More in general, various studies, especially those on well-being, have aimed to analyse how "happy" people were. Also, within international organizations the importance of this feature has been widely recognized leading to the inclusion of questions about "happiness" in various studies including the Eurobarometer and World Values surveys. Yet, these data are limited both in time and geographical coverage. One exception is Hills et al (2015), who used a dataset from google books to make some tentative estimates of happiness for the past two centuries for a few European countries.

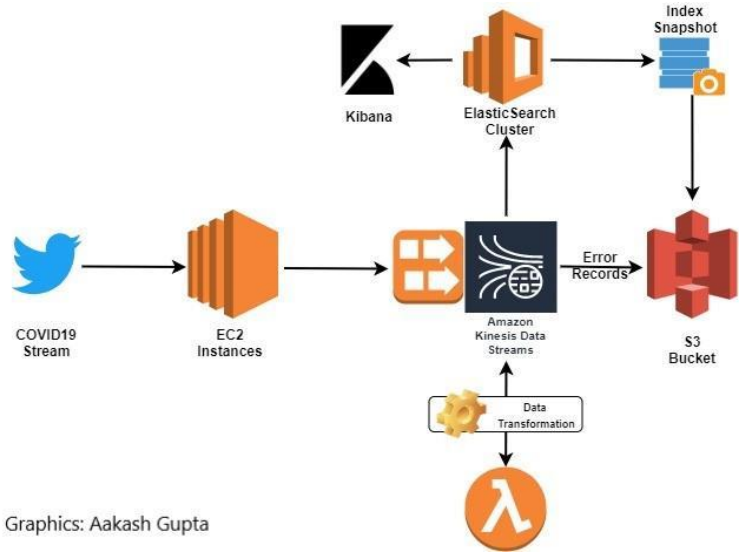
Unfortunately, due to lack of data mining tools, this field is still in its infancy, which meant that, for example, the authors of the recently published OECD "How Was Life?" report, which covered various aspects of well-being over the past two centuries, had only indirect indicators used to capture societal sentiments. Besides this lack of studies on happiness, even fewer tools are available that move away from "general happiness" to topic related sentiments such as how people feel about topics such as migrants, monetary expansion, etcetera. A clear example of the importance of topic related sentiment is the recent Covid pandemic, which opened up a debated on the valuation of "crucial" occupations such as hospital workers.

In this paper we make an attempt to apply these questions related to covid to a large dataset of tweets in the Netherlands between January and September 2020 (Section 2). In using these data, we check the number of Covid-related tweets by region (section 3) and how people feel (the sentiment score, see Section 4), as well as offering a preliminary analysis of how this affected the valuation of occupations (Section 5). In Section 6, we end with a brief conclusion.

2. Twitter Dataset

Through the Twitter developer labs, we have a streaming end-point for all Covid19 dialogues. It delivers free, full-fidelity data in real-time on the COVID-19 conversation. Even though it's just 1% of the conversation it's still a good sample of the social dialogue. The data ingestion architecture is illustrated in Figure 1. The Twitter API endpoint is called for hydrating the tweet IDs. The returned data has more than 250+ data points.

Figure 1. Cloud Architecture for hydrating & filtering the tweets



Since we focus on the Netherlands, we use the following filter to select the tweets of interest:

1. Tweet language == Dutch; OR
2. Tweet user-location ~ Netherlands

The hydrated tweet is then put into a data stream which writes into a Kinesis Firehose. Lambda function is used to process the data in batches. We use this to select the required keys, process the tweet text and store it in an AWS managed ElasticSearch index. AWS S3 bucket is used for backup & storing failed batches.

The resulting data still requires expansion with province, industry classification of each tweet, and sentiment score. First, the tweets do not report province but rather exact user location. Hence, we use OpenStreetMaps for the geo-coding of users-location data.[1] Nominatim API is used with a user_agent for geocoding the user location. Nominatim indexes named (or numbered) features within the OpenStreetMap (OSM) dataset and a subset of other unnamed features (pubs, hotels, churches, etc). This latter data is sparse as only *89 unique point-information* is extracted from the tweets. This data is then mapped with provincial shape files to identify if the user-location lies within its geographical boundaries. Out of the 271,342 tweets, we were able to record 107,127 tweets for the province. Even though they might still

[1] The OpenStreetMap Foundation is an international not-for-profit organization supporting, but not controlling, the OpenStreetMap Project. It is dedicated to encouraging the growth, development and distribution of free geospatial data and to providing geospatial data for anyone to use and share.

contain biases, this bias is probably less since we record provinces rather than individual cities. In addition, Sloan et al (2013) argue that the geo-located observations move in line with population totals.

Second, to find industries in the tweets, we used text classification for HISCO codes. As a proof-of-concept we have built a data pipeline, which iterates through the tweet text and identifies words which are similar to HISCO standard keywords. This is then used to provide an industry classification.

Figure 2. identification HISCO codes in tweets

Tweet:

Leraar middelbare school besmet. Tien andere leerkrachten zijn uit voorzorg thuis. Zij voelen zich ook ziek. Dit verspreidt als een malle😫
#Coronanederland #CoronavirusNederland #CoronavirusNL #Covid19NL

Entity Recognition:

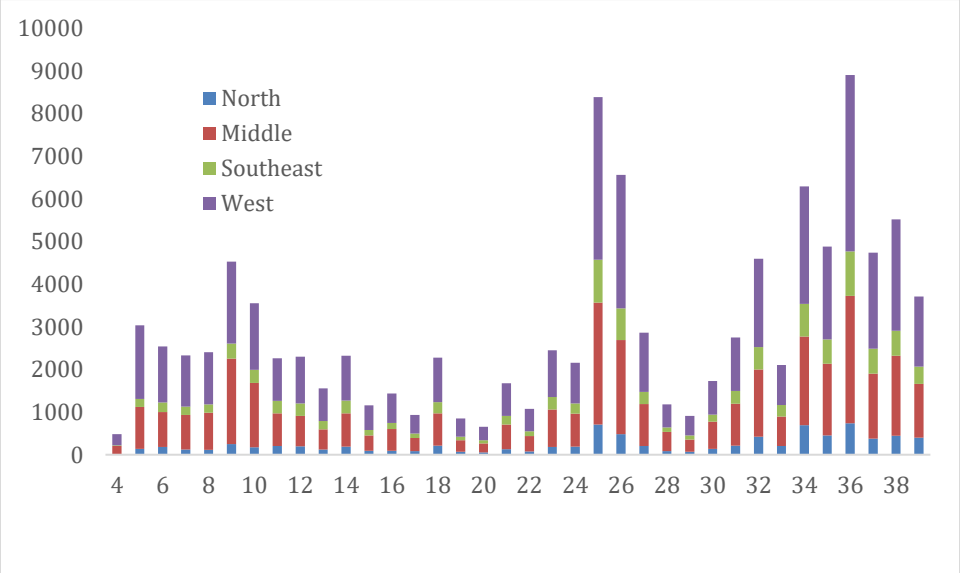
Leraar middelbare school besmet. Tien andere leerkrachten zijn uit voorzorg thuis. Zij voelen zich ook ziek. Dit verspreidt als een malle😫

Third, in order to attach a sentiment analysis to each tweet, we apply a sentiment analysis, which recently became quite popular in the literature (e.g. Chakraborty 2020; Manguri 2020; Mansoor 2020). It allows processing of all available messages and data and extracting their sentiments and indicating if they're positive, negative or neutral. Python's Pattern library (https://github.com/clips/pattern) offers a useful Natural Language Processing library and has Data Mining web services for Google, Twitter and Wikipedia. This library has a rich functionality and is able to perform tasks such as tokenization (pluralizing and singularizing), stemming, part of speech tagging, finding N-grams, etc. Regarding sentiment analysis, Pattern library is assigning scores called the polarity to positive (good, best, excellent, etc.) and negative (bad, awful, pathetic, etc.) adjectives, a sentiment score between 1 and -1 is assigned to the text. Every message from Twitter got this sentiment score in the final report, neutral messages got 0.

3. How about Covid?

Of the total number of Covid tweets in our dataset (271,327), those that allow identification by province (107,123) are reported by week in Figure 3. As can be seen, tweets are unequally divided over provinces with the West (North and South Holland and Zeeland) and Southeast (Gelderland, Limburg, and North-Brabant) having more tweets than the Middle (Utrecht and Flevoland) and North (Friesland, Groningen, Drenthe, Overijssel), arguably because of their larger population shares.

Figure 3. No. of Covid tweets by week of 2020 and region in the Netherlands



Since these numbers may simply reflect the popularity of twitter in various regions, these absolute numbers tell us little. Yet, one thing we can conclude is that there is little evidence of different trends in number of tweets over time among regions: we find peaks in tweets in all regions around week 9-10 (17 February -8 March), week 25-26 (15 June-28 June), and week 34-38 (17 August – 20 September).

4. Sentiments

The peak in tweets that run from 17 February to 8 March, coincides with the rise in number of infections. Likewise, the peak at week 34-38 (August/September) coincides with the rise towards the October peak (Figure 4).[2] The exception is week 25-26 (end of June), which does

[2] https://www.rivm.nl/coronavirus-covid-19/grafieken

not coincide with a rise in infections but is probably related to the emergency law ("Spoedwet", het wetsvoorstel *Tijdelijke wet maatregelen COVID-19*) that allowed the government to curb certain freedoms to fight the spread of Covid and was strongly opposed by certain groups in society. This increased the number of tweets. Indeed, tweets are often found to be related to policy measures (Kruspe et al. 2020; Daas 2020, 6). Hence, with the exception of June, the number of over-all infections lag by about a month to the number of tweets as can be seen in Table 1 where the number of infections lagging by about a month (month *t-1*) are positively and significantly correlated to the number of tweets in month *t*.

Figure 4. No. of Covid tweets and infections by week of 2020

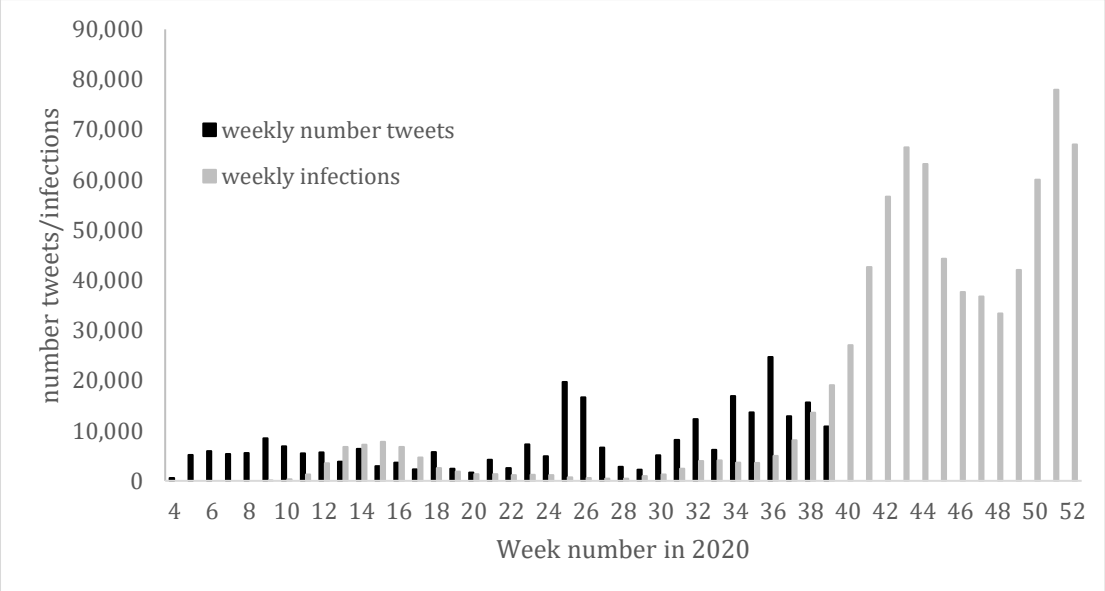


Table 1. Correlation number of tweets by region and number of infections

		no. tweets			
		North	Middle	Southeast	West
no. infections	current infections	0.2063	0.0811	0.1388	0.1311
	Infections -1 month	0.4849***	0.04009***	0.3959***	0.3473**
	infections -2 months	-0.0888	-0.0560	-0.1184	-0.0307

*20% significance, **10% significance, ***5% significance

One explanation of this lag in infections compared to the number of tweets might be the feeling of people about Covid. Just as other studies (e.g. Bhat et al 2020), we find Covid tweets to be mainly positive or neutral (see Figure 5). These positive sentiments are related to

the number infections though: during peaks (e.g. October) sentiment scores were highest (more positive), which led to lower numbers of tweets. On the contrary, sentiment scores were lowest (more negative) during periods with few infections, leading to more tweets. This implies a positive correlation between infections and sentiments and a negative or insignificant one between tweets and sentiments (see Table 2). Our previous observation that the peak in number of tweets in June was unrelated to the number of infections is also shown by low sentiment scores in July, probably caused by the "spoedwet" and the, sometimes violent, protests against it.

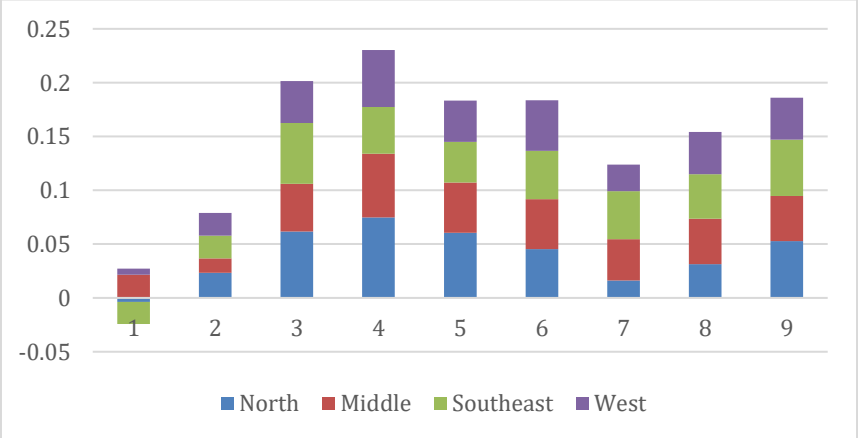
Table 2. Correlation by weekly mean sentiment by number of infections & number of tweets

	Sentiment score			
	North	Southeast	Middle	West
no. of infections	0.2961*	0.3852***	0.2433*	0.1925
no. of tweets	-0.0651	0.1188	-0.0211	0.1816

*20% significance, **10% significance, ***5% significance

It is important to note that, even though overall patterns of increases in sentiments during peaks of infections are similar among regions, the North and West regions show biggest drops in positive sentiments in both February-March and July (and biggest rises thereafter) (Figure 5).

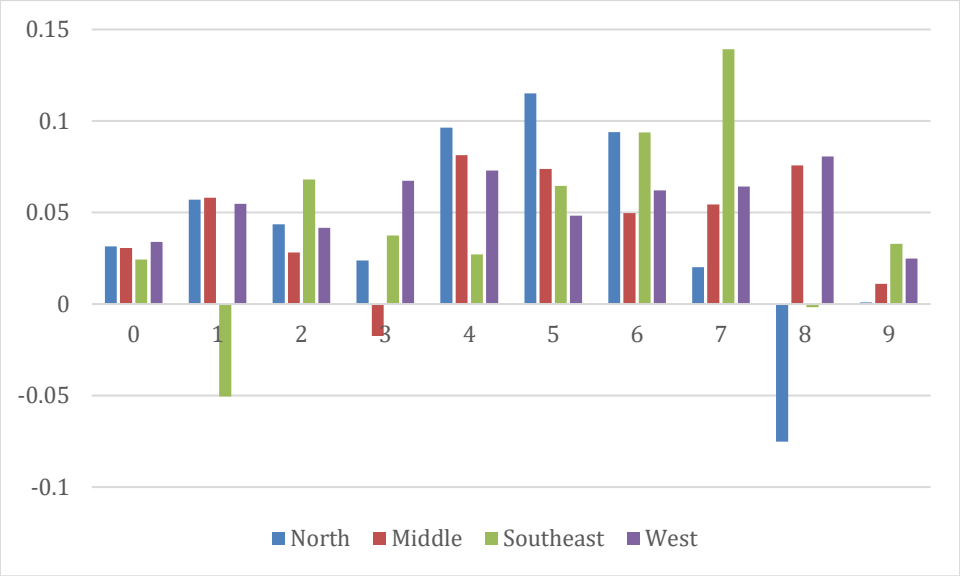
Figure 5. Cumulative sentiment score of Covid tweets by month of 2020 and region in the Netherlands



5. Valuation occupations

The reason for declining number of tweets during peak infections thus appears to be the more positive attitude during those times. This may be due to various reasons. One reason of the increase in positive sentiments during peak infections might be how people value societal response, more specifically the "crucial" occupations. We plot in Figure 6 the mean sentiment score of tweets mentioning a 1-digit HISCO occupation. Group 0-3, 9 (i.a. teachers and ordinary labourers), arguably comprising most of the "crucial" occupations, are mentioned in the least positive context, while groups 4-8 (a.o. clerical, sales, service, and agricultural workers) were valued more positively.

Figure 6. Mean sentiment score of Covid tweets by 1-digit occupation by region in the Netherlands



Yet, even though sentiment scores of crucial occupations were lower, they did increase in periods of peak infections while non-crucial occupations did not (Table 4). One might argue that this distinction in sentiment scores for "crucial" versus other occupations is due to subjectivity (Table 3): higher social positions for crucial occupations corresponds with lower subjectivity, that is more objective visions on crucial occupations during peak infections. Yet, after the peak is over, these crucial occupations return once more to their lower social status with its associated increased twitter attention.

Table 3. Correlation of sentiments, subjectivity, and social position

month	sentiments*subjectivity	sentiments*social position	subjectivity*social position
February	0.445***	-0.135	-0.473***
March	-0.111	-0.218	-0.087
April	-0.063	-0.088	-0.328***
May	-0.105	0.105	0.072
June	-0.062	0.105	-0.389***
July	0.000	0.081	-0.085
August	0.154	-0.135	0.134
September	0.211	-0.278	-0.269**

*20% significance, **10% significance, ***5% significance

Table 4. correlation weekly infection with mean sentiments by HISCO groups 0-3/9 and 4-8.

	Crucial occupations	non-crucial occupations
no. infections	0.2516*	-0.1071

*20% significance, **10% significance, ***5% significance

6. Conclusion

We use numbers of Covid-related tweets by Dutch region in month 1-9 of 2020. Peaks in tweets precede the number of infections by about one month (exception is June when the "Spoedwet" was introduced). The reason is that positive sentiments occurred in peak infections leading to fewer tweets, while in periods of rising infections more negative sentiments dominated that were voiced in an increased number of tweets.

This increased positivity during peak infections may have various reasons. Yet, for one, it may be due to how persons value society. We find, a higher valuation of "crucial" occupations during the peaks in infections, which is combined with fewer twitter mentions. Indeed, the higher the number of infections, the more positive the sentiments related to crucial occupations, while this relation does not hold for non-crucial occupations.

References

The twitter dataset is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License (CC BY-NC-SA 4.0). By using this dataset, you agree to abide by the stipulations in the license, remain in compliance with Twitter's *Terms of Service*, and cite the following manuscript: Chen E, Lerman K, Ferrara, E. (2020). "Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set." *JMIR Public Health Surveill* ;6(2):e19273 DOI: 10.2196/19273 PMID: 32427106

Bhat, M., M. Qadri, N. Beg, M. Kundroo, N. Ahanger, and B. Agarwal (2020). "Sentiment analysis of social media response on the Covid19 outbreak." *Brain, behavior, and immunity* 87: 136-137. doi:10.1016/j.bbi.2020.05.006

Chakraborty, K., S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. Hassanien (2020). "Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media." *Applied Soft Computing* 97, Part A: 106754.

Daas, P.J.H. (2020). "Ontwikkeling van een Corona sentimentsindicator:Methodologische verantwoording." *CBS*. Downloaded from https://www.cbs.nl/nl-nl/over-ons/innovatie/project/corona-sentimentsindicator

Hills, Th., E. Proto, and D. Sgroi (2015). "Historical analysis of national subjective well-being using millions of digitized books," IZA Discussion Paper No. 9195.

Kruspe, A., M. Haberle, I. Kuhn, and X.X. Zhu (2020). "Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic." ACL Anthology. https://www.aclweb.org/anthology/2020.nlpcovid19-acl.14.pdf

Manguri, K. H., R. N. Ramadhan, and P. R. Mohammed Amin (2020). "Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks." *Kurdistan Journal of Applied Research* 5 (3): 54-65.

Mansoor, M., K. Gurumurthy, R.U. Anantharam, and V. R. Badri Prasad (2020). "Global Sentiment Analysis Of COVID-19 Tweets Over Time." arXiv:2010.14234v2

OpenStreetMap database is used for geocoding, licensed under the Open Data Commons Open Database License (ODbL) by the OpenStreetMap Foundation (OSMF).

Sloan, L., J. Morgan, W. Housley, M. Williams, A. Edwards, P. Burnap, and O. Rana (2013). "Knowing the tweeters: Deriving sociologically relevant demographics from twitter." *Sociological Research Online* 18(3): 7.