


## Annex 1. Tweet geolocation

Our first dataset candidate contained all tweets in 2021 that contained at least one of selected VACCINE\_KEYWORDS and at least one of selected GEO\_KEYWORDS.

The VACCINE\_KEYWORDS are chosen as: "vaccine", "vaccination", "vax", "antivax", "anti-vax", "anti-vaccine", "antivaccine", "vaxed", "vaxxed", "unvaxed", "unvaxxed", "vaccinated", 

The GEO\_KEYWORDS were chosen using a multi-step procedure, in which we geolocate South African twitter users, extract their followers, and extract keywords (words, hashtags, ngrams, mentions) from their vaccination-related tweets which are a) used frequently and b) clearly reference a South African context. The details are laid out further below.

1. We search original tweets for VACCINE\_KEYWORDS in 2021 using the operator `geodata_country = "ZA"` and extract authors of tweets. Authors were then double checked by searching for reference to South Africa (including the country flag), and South African towns and provinces in the profile's description and location. These users were hence filtered by either being a news source or having at least 100'000 followers. These will be "seed authors" (N = 1137) that are clearly associated with SA and tweet about vaccines.
2. We get the subset of "followers" of the seed authors which follow at least 50 of them, to widen the sample. The threshold of 50 was established by hand-checking a sample of 160 users profiles, stratified by the number of followed seed authors. Up to 20 followed seeds there was still a high probability (~50%) that the user was not in or connected to South Africa. This probability dropped remarkably (~20%) for users following 50 to 70 SA seeds. The final followers sample had N=630'183
3. We search original tweets for VACCINE\_KEYWORDS in 2021 from the 630'183 followers and the 1137 seed users
4. The resulting set of seed author tweets (n = 86'337) + follower tweets (n = 1'675'891) is cleaned, and n-grams are formed.
5. From the resulting text corpus, all n-grams, mentions and hashtags are extracted and sorted by their frequency in the corpus.
6. A selection of local "tags" based on the most frequent n-grams, mentions, hashtags was compiled by hand. Here some freedom of judgement was used, for example n-grams "health\_minister\_dr\_zweli" and "dr\_zweli\_mkhize" were merged to "zweli\_mkhize", or ambiguous hashtags like #vaccination were excluded even if they relate to a South African vaccine campaign.
7. This final subset is used as GEO\_KEYWORDS: south africa, south africans, south african, ramaphosa, south africas, western cape, news24, julius malema, david makhura, vaccinerolloutsa, cape town, nicd sa, discovery sa, richard spoor, presidencyza, zweli mkhize, eastern cape, gautenghealth, mbusi b ndlovu, jj stellies, thecitizen news, david mabuza, ms zamandlovu, joe phaahla, sisonke trial, jacob zuma, magda wierzycka, unathi kwaza, clicks sa, bhekisisa mg, #vaccinerolloutsa, #covid19sa, #familymeeting, #sabcnews, #vaccineforsouthafrica, #vaccinatetosavesouthafrica, #vaccinerolloutsa, #ichoosesevaccination, #covid19insa, #enca, #southafrica, #dstv403, #ramaphosa, #cyrilramaphosa, #voomavaccination, #newzroom405, #ramaphosamustfall, #shutdownsa, #unityinaction, #juliusmalema, #sisonke, #voetsekanc, #thankyoueff, #sona2021, #alcoholban, #thumaminamedia group, #smile904fmnews, #capetown, #zwelimkhize, #safmsunrise, @cyrilramaphosa, @news24, @healthza, @miamalan, @drzwelimkhize, @enca, @sizwelo, @heidigiokos, @sentletse, @mzwanelemanyi, @gautenghealth, @bisouthafrica, @governmentza, @sabcnews, @presidencyza, @therealpro7, @timeslive, @abramjee, @vngalwana, @safmnews, @mithisa\_motho, @gautengprovince, @danielmarven, @moshabelamosa, @geoffreyyork, @cmfundisi, @ewnupdates, @iol, @advobarryroux, @karynmaughan

---

## Annex 2. Support Vector Machine detailed results

### *Confusion Matrix*

	Provax	Antivax	Unclear	Facts
Provax	82	8	23	2
Antivax	24	52	24	1
Unclear	6	5	71	5
Facts	5	0	5	133

### *Detailed statistics*

	Precision	Recall	F1-score
Provax	0.71	0.7	0.71
Antivax	0.51	0.8	0.63
Unclear	0.82	0.58	0.68
Facts	0.93	0.94	0.94
Accuracy			0.76
Macro Avg	0.74	0.76	0.74
Weighted Avg	0.78	0.76	0.76