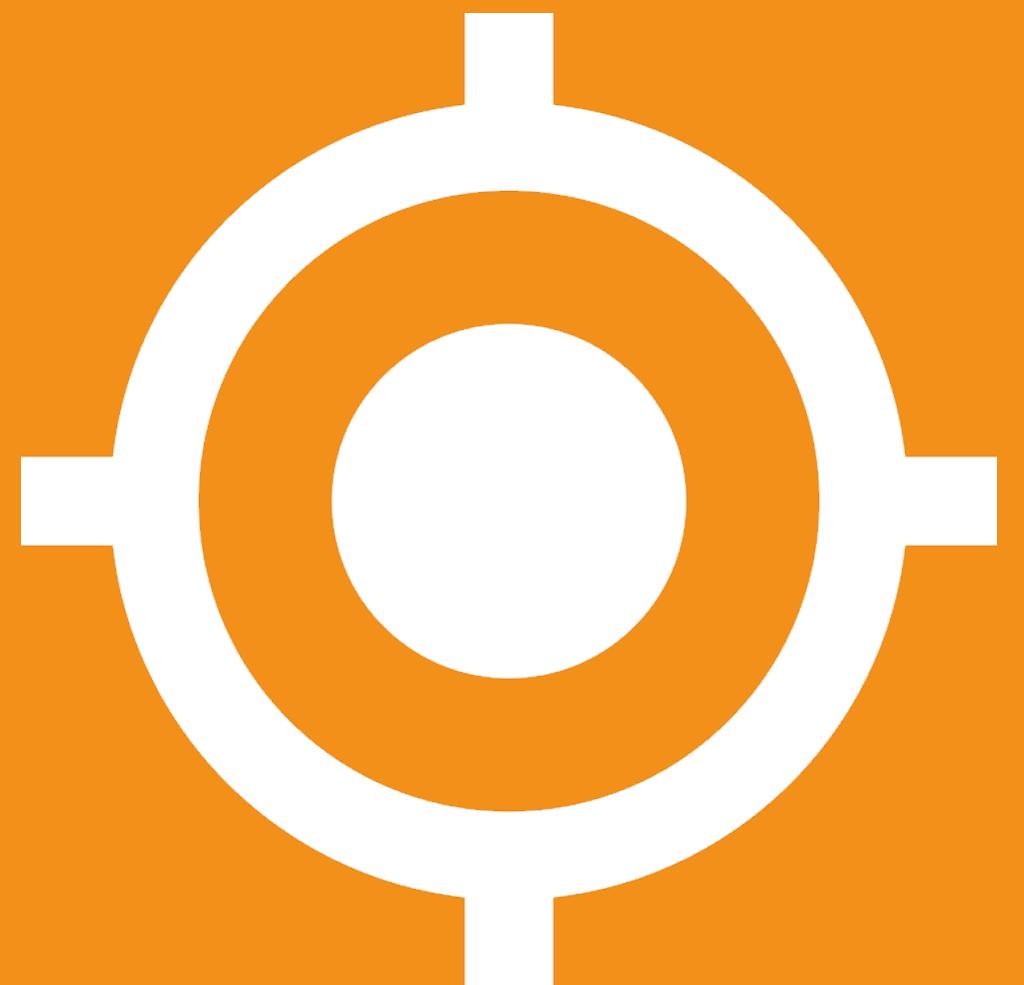


Location, location, location!

Locatiebepaling van Nederlandse tweets via tekstclassificatie

Léon Melein

Reinard van Dalen



Onderzoek

1. Onderzoeksvorag
2. Materiaal
3. Methode en Evaluatie
4. Resultaten
5. Discussie
6. Conclusie





Onderzoeksvraag: onderwerp

*"Wij onderzoeken de toepasbaarheid van
de Naive Bayes-classificatiemethode voor
het classificeren van Nederlandstalige
twitterberichten op basis van de provincie
waaruit ze afkomstig zijn,*

Onderzoeksvraag: vraagstelling

omdat wij willen weten in hoeverre deze classificatiemethode de afzendlocatie van een bericht kan herleiden uit de inhoud hiervan,

Onderzoeksvraag: doelstelling

teneinde te kunnen beoordelen of deze methode kan helpen bij de locatiebepaling van twitterberichten."

Materiaal

- Nederlandstalige tweets uit augustus 2015
- Verzameld door de Rijksuniversiteit Groningen
- Filtering op twee criteria:
 - Tweets met geotags
 - Tweets uit Nederland



```
{  
    "created_at": "Mon Jan 04 10:21:34 +0000 2016",  
    "id": 683956427828912100,  
    "text": "Rond Zuidhorn is de dans van de hoogspanningslijnen inmiddels voorbij.  
        #horrorwinter #lijndansen #stroomdippen",  
    (...)  
    "geo": null,  
    "coordinates": null,  
    "place": {  
        "id": "182e62b1b1cccd2b3",  
        "url": "https://api.twitter.com/1.1/geo/id/182e62b1b1cccd2b3.json",  
        "place_type": "city",  
        "name": "Zuidhorn",  
        "full_name": "Zuidhorn, Groningen",  
        "country_code": "NL",  
        "country": "Nederland",  
        "contained_within": [ ],  
        "bounding_box": (...),  
        "attributes": {}  
    },  
    (...)  
    "lang": "nl"  
}
```

Materiaal

Verzamelen Nederlandstalige tweets

```
zcat /net/corpora/twitter2/Tweets/2015/08/* | /net/corpora/  
twitter2/tools/tweet2tab id user text place | wc -l
```

- 20.372.259 tweets

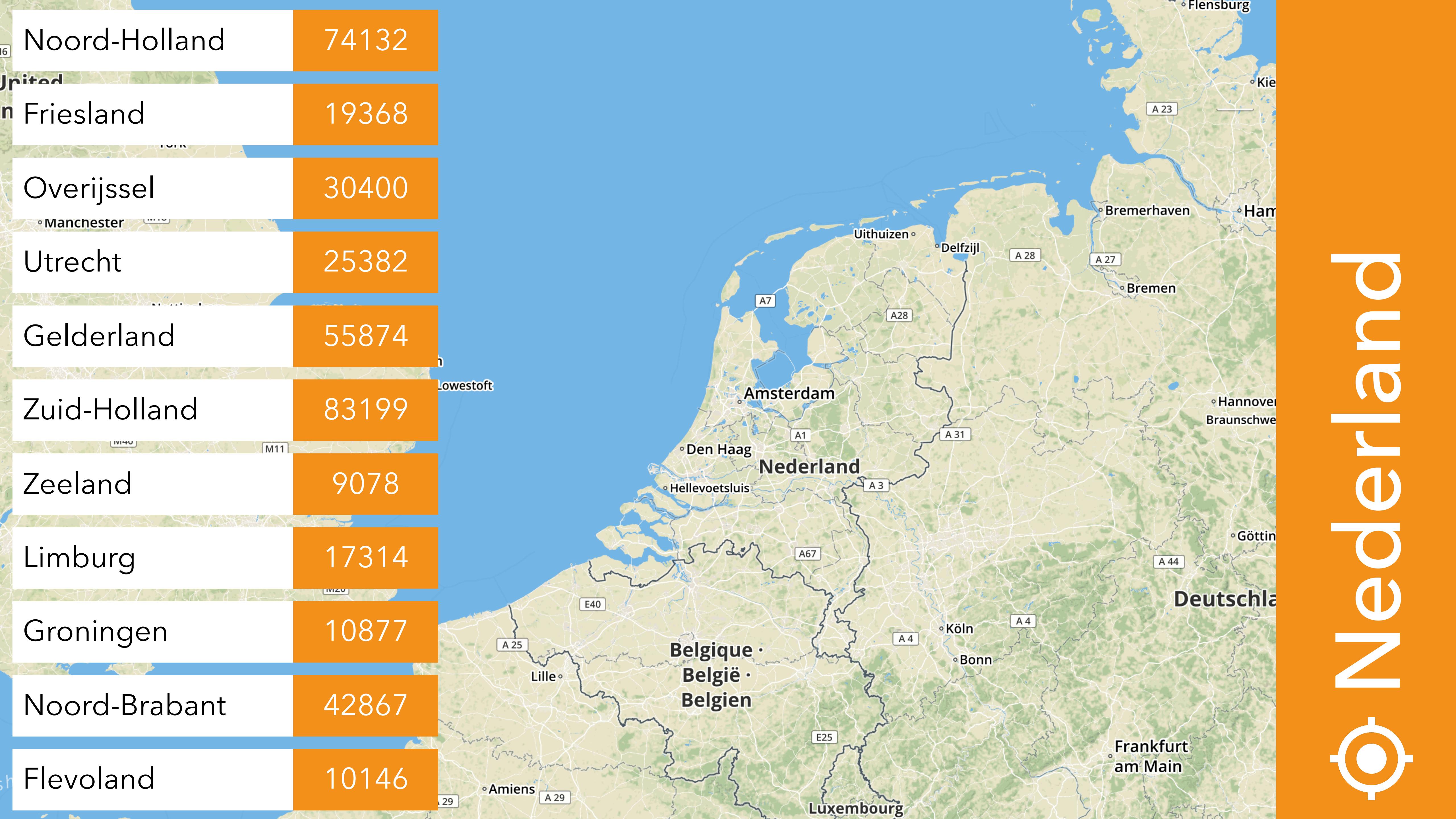
Materiaal

Verwerking tot corpus

- Tweets onderverdeeld in mappen per provincie
- Python + Pickle
- 378.637 tweets
- 90% trainingsdata, 10% testdata

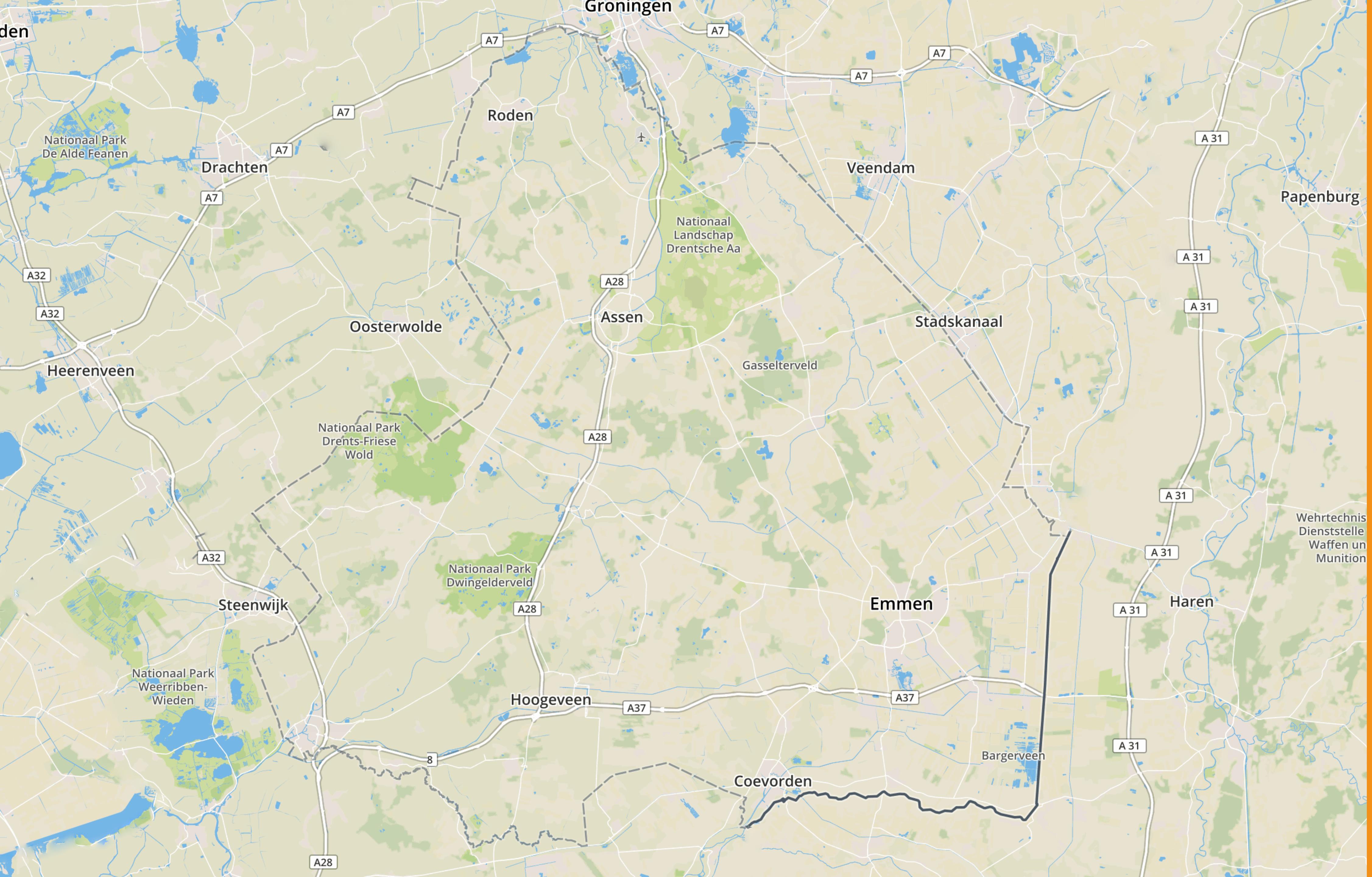
```
{Friesland: [tweet_1, tweet_2, ...],  
Groningen: [tweet_1, tweet_2, ...],  
...}
```

Nederland

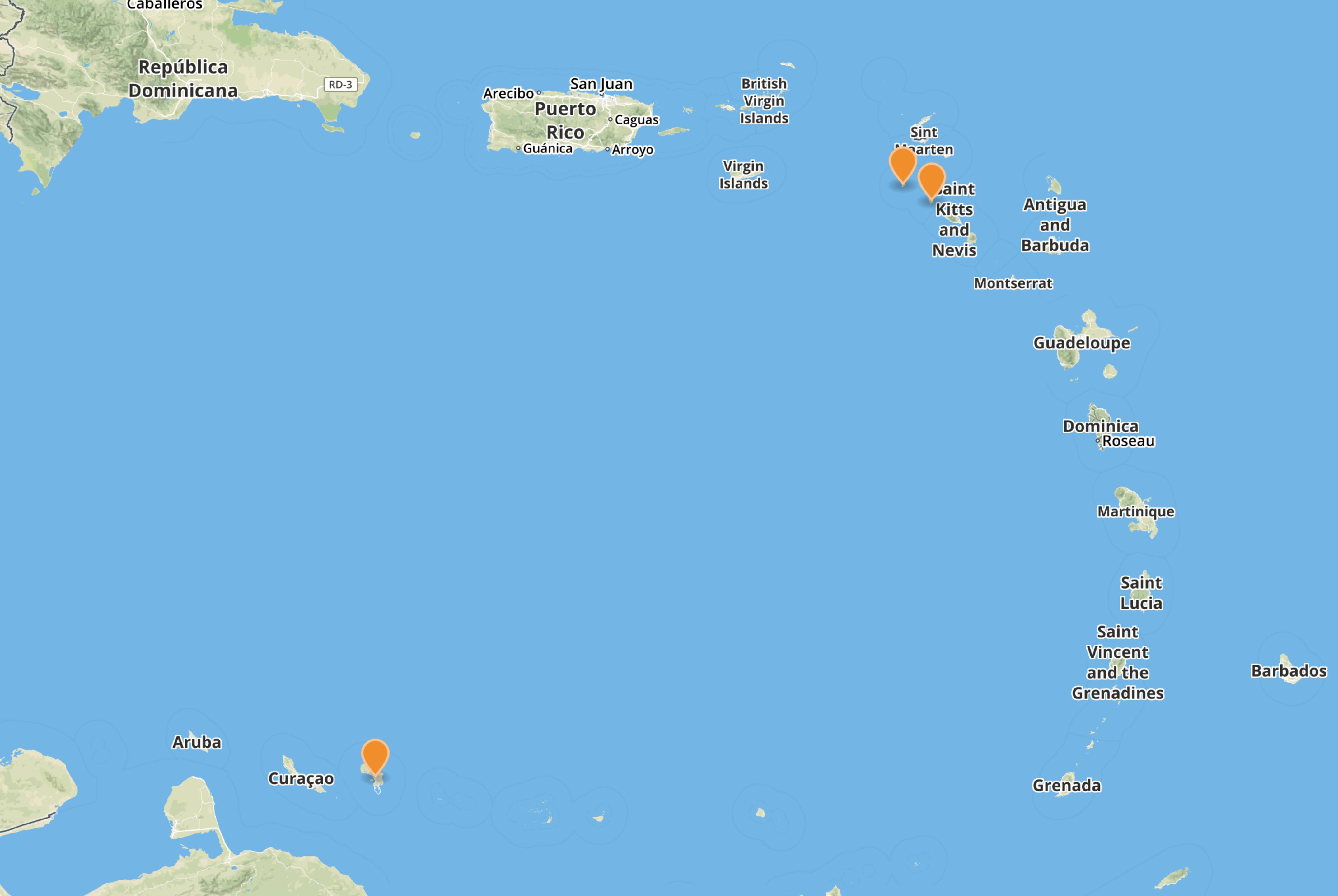
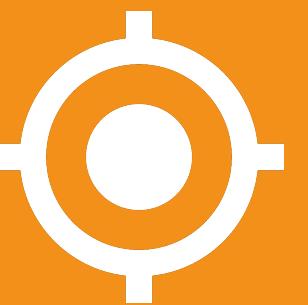
A map of the Netherlands with provincial boundaries. Major cities like Amsterdam, Rotterdam, and Utrecht are labeled. Neighboring countries shown are Germany (Deutschland), Belgium (Belgique - België - Belgien), and Luxembourg. A road network is also depicted.

Drenthe?

Drenthe?



Drenthe?



Bonaire,
Sint Eustatius & Saba
(\pm 7915 km)

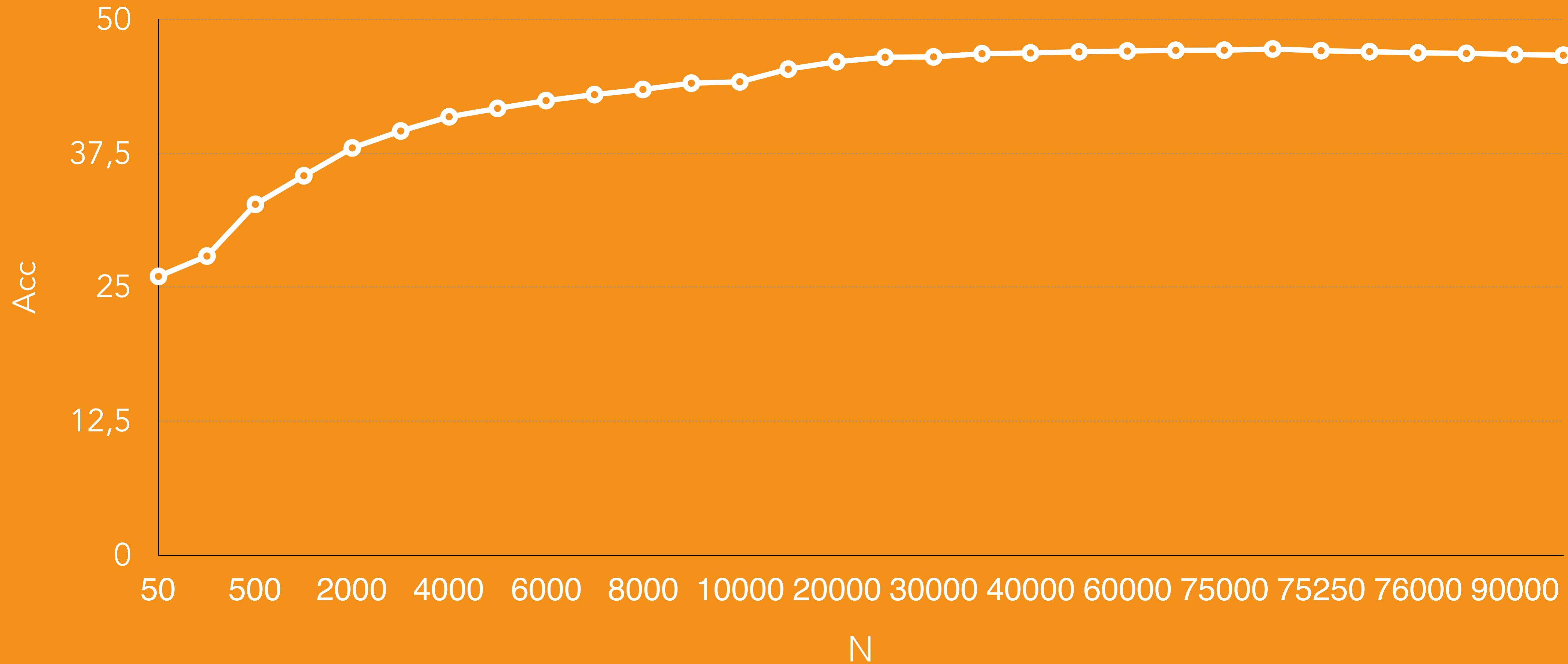
Methode en evaluatie

- *Rainbow classificatietool*
- 10-fold cross validation
- Baseline meting met *Naive Bayes*
- Effect van *feature selection*
- Evaluatie op basis van Accuracy

Resultaten: baseline

- Accuracy: 44,4%
- Precision: 30,0%
- Recall: 66,0%
- F1-score: 41,3%

Resultaten: feature selection



Resultaten: feature selection

$N = 75125$

- Accuracy: 47,27%
- Precision: 35,65%
- Recall: 64,87%
- F1-score: 45,99%

Discussie

- Beperkingen dataset
 - Uitsluitend tweets binnen Nederland
 - De provincie Drenthe ontbreekt
 - Slechts één maand data (378.637 tweets)

Discussie

- Beperkingen methode
 - Uitsluitend contentgebaseerd
 - Gebrek aan locatiespecifieke informatie

Conclusie

Baseline

- Accuracy: 44,4%
- Precision: 30,0%
- Recall: 66,0%
- F1-score: 41,3%

Feature selection

- Accuracy: 47,27%
- Precision: 35,65%
- Recall: 64,87%
- F1-score: 45,99%

Conclusie

- Geen betrouwbare classificatie
 - Beperkte dataset
 - Uitsluitend contentgebaseerde methode
- Verder onderzoek blijft noodzakelijk