STATISTIEK VLAANDEREN

National Sentiment Statistics through Social Media: Obstacles and Opportunities

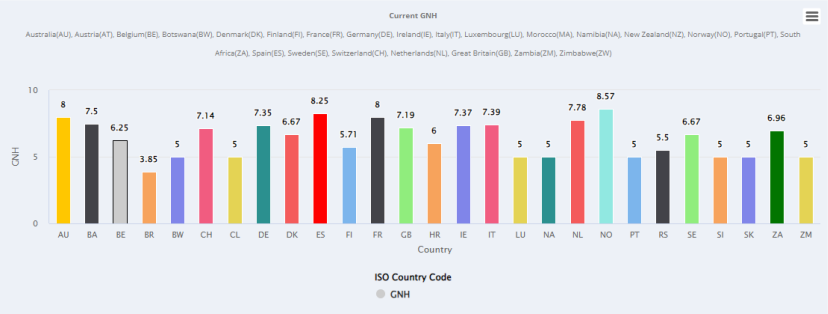
Michael Reusens
Joint work with Cedric De Boom, Manon Reusens et al.
Statistics Flanders – Belgium
Statistics Flanders Data Science Hub

13-03-2024
EMOS Seminar

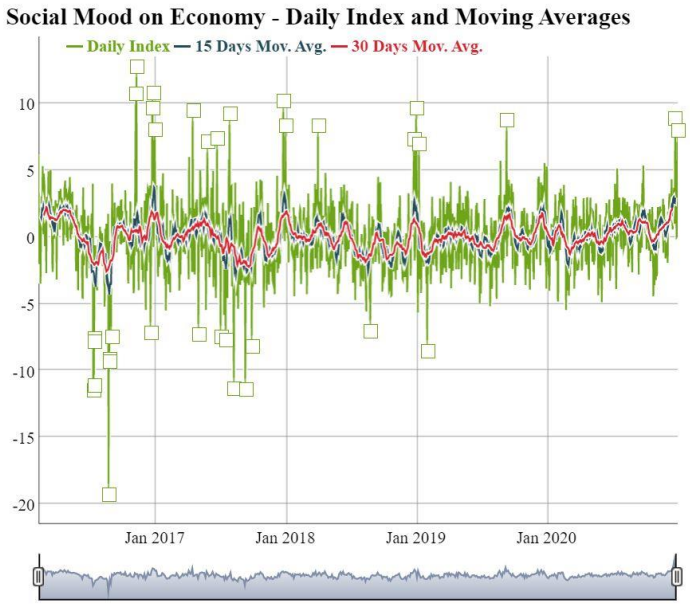
Vlaanderen
is sterk in cijfers

Inspired by others

GNH.today

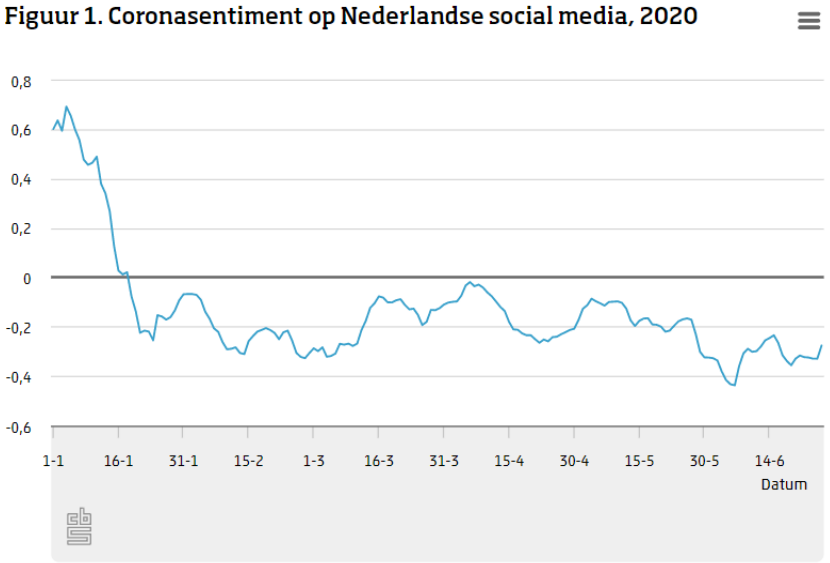


Italy: social mood on economy



https://www.istat.it/en/archivio/219600

Netherlands: corona



https://www.cbs.nl/nl-nl/over-ons/onderzoek-en-innovatie/project/corona-sentimentsindicator

STATISTIEK VLAANDEREN

Vlaanderen is sterk in cijfers

The Flemish Twitter Sentiment

= a statistic based on the sentiment of Flemish Twitter users

Sentiment?

= the im-/explicit opinion of the author about the subject of their tweet

Antwerp FTW!

Hudson water level has risen

Down with the flu, again… ☹

Why Twitter/social media?

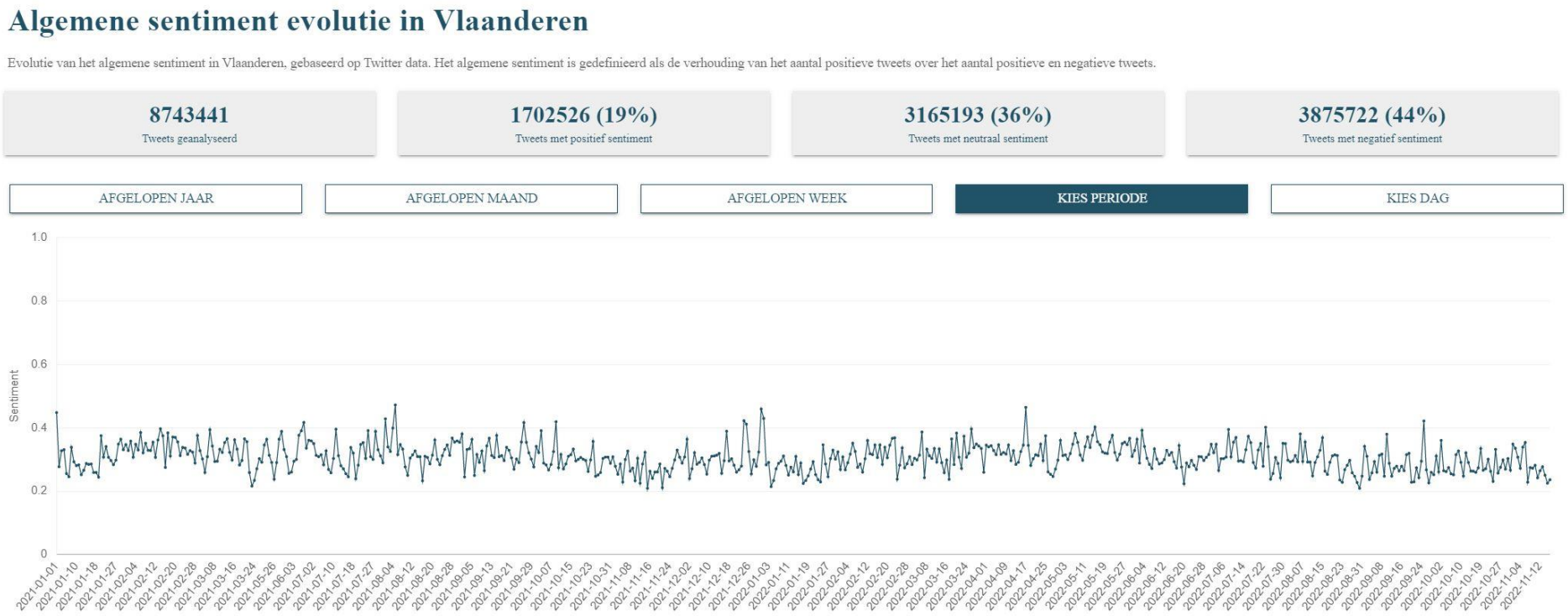
- High frequency
- Quasi realtime
- Open platforms

A proverbial “finger on the pulse”

- More generally
  - Increasing demands on more and better statistics
  - Innovate to make sure official statistics stay relevant!

What does it look like

Screenshot of our internal testing dashboard



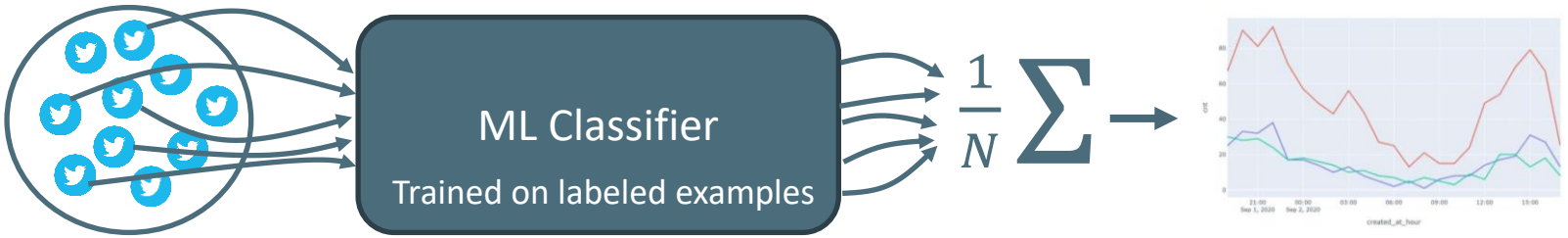
STATISTIEK VLAANDEREN

Vlaanderen is sterk in cijfers

The birds-eye view (seems easy!)

Classification:

Aggregation:



Some important questions (pandora's box)

Data

- How to obtain the data, automatically?
- How to obtain labels?
- Ambiguous labels?
- Annotator bias?
- Data source dependencies?

Model

- What kind of predictive model?
- Model staleness? Concept drift? Retraining?

Statistic

- How to aggregate model predictions?
- How to provide insights / explainability?
- Selection bias?
- Good model = good statistic?

Collaboration with academia

- **Comparison of Different Modeling Techniques for Flemish Twitter Sentiment Analysis**. Analytics. 2022, Reusens M, et al.

- **Predicting annotation difficulty to mitigate annotator bias,** Master Thesis, 2022, De Waele S., Vanderschueren M., Bache-Mathiesen J., et al.

- **Predicting the demographics of Twitter users with programmatic weak supervision**. *TOP*, 2024, 1-37, Tonglet, J., Jehoul, A., et al.

- **Quality of life in flanders: a comparative study using twitter and survey data.** Master thesis, 2023, Vranken S., Ferket N.,et al.

- **Measuring political confidence using twitter sentiment analysis: a belgian example.** Master thesis, 2023, Van Poppel O., Chi Chung Fong M., et al.

- **Adding context to the Flemish Twitter sentiment indicator using natural language processing**. Master thesis, 2023, Randriamanana N., et al.

Which data?

Live filtering through Twitter API v2

Twitter filters 1% of all tweets.

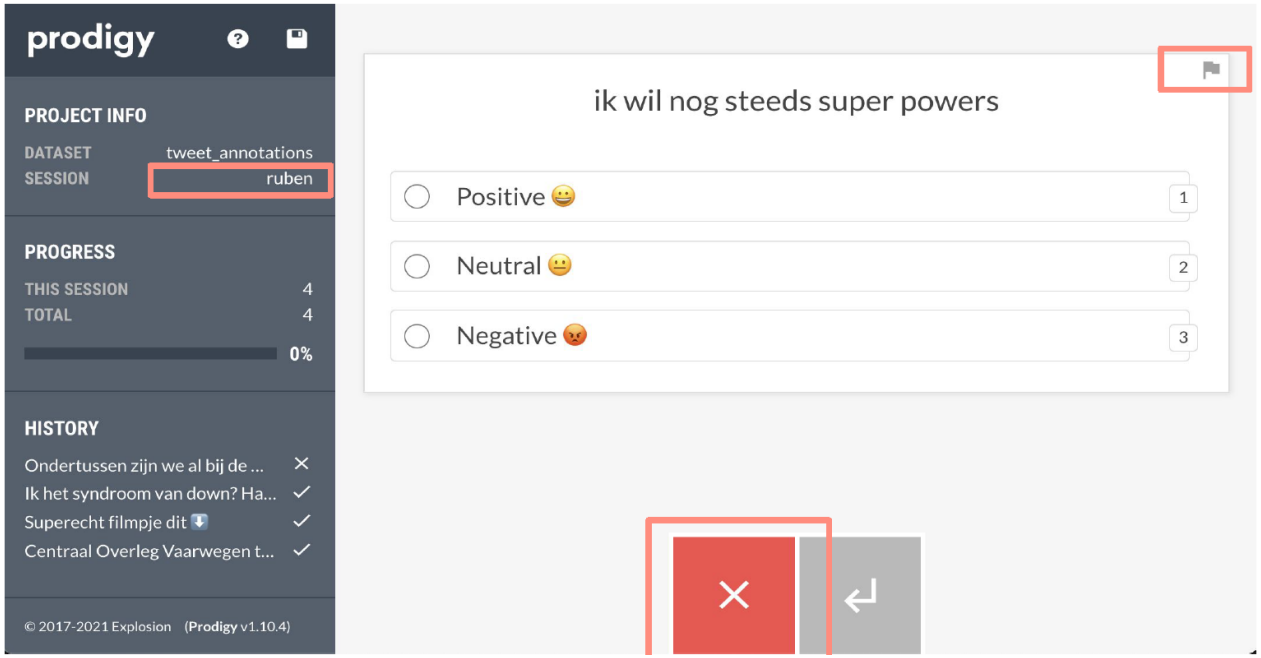
We implement extra filters:

- ✓ Dutch language
- ✓ From Belgium, Flanders or a Flemish town/city
- ✘ No retweets nor replies
- ✘ No media (photos, videos…)

Labelling in practice

3 annotators annotated **50K tweets** between Oct 2015 and Oct 2020

Part of the dataset was annotated by multiple annotators
in order to be able to inspect differences in annotation



STATISTIEK VLAANDEREN

Vlaanderen is sterk in cijfers

Labelling: the manual

1. **Question:** is the tweet in proper Dutch and do you understand what it says?
   a. **Yes:** continue to (2).
   b. **No:** discard the tweet.
2. **Question:** Does the tweet contain explicit information that reveals the author's sentiment directly? Examples: author states his/her opinion, emoji reveal sentiment, hashtags reveal sentiment.
   a. **Yes:** label the tweet according to the revealed sentiment.
   b. **No:** go to (3).
3. **Question:** Does the tweet resemble a factual statement?
   a. **Yes:** Is the factual statement personal to the author and at the same time not personal to the general public? Examples: a fact about the author or his/her relative, his/her belongings, his/her immediate neighbourhood.
      i. **Yes:** Is there exactly one clear sentiment the author could experience with respect to this factual statement?
         1. **Yes:** label the tweet according to how the author would experience the factual statement.
         2. **No:** flag the tweet since it is ambiguous which sentiment to use (either because there are multiple valid answers or because it is unknown which sentiment the author would experience).
      ii. **No:** label the tweet as *neutral*.
   b. **No:** go to (4).
4. **Question:** Does the text have an implicit sentiment from which you could infer the author's intended sentiment? Example: sarcastic question
   a. **Yes:** Is there exactly one clear sentiment the author could have intended implicitly?
      i. **Yes:** label the tweet according to the implicit sentiment.
      ii. **No:** label & flag the tweet since it is ambiguous which sentiment to use (either because there are multiple valid answers or because it is unknown which sentiment the author intended).
   b. **No:** label & flag the tweet.

Improving labelling efficiency

We used two methods to improve labelling efficiency:

Choose tweets that are sufficiently different from each other

1. Compile a list of >200K tweets.
2. Encoded every tweet using a multilingual sentence representation.
3. Compare the tweets against each other.
4. Use only the “most unique” tweets.

Active learning

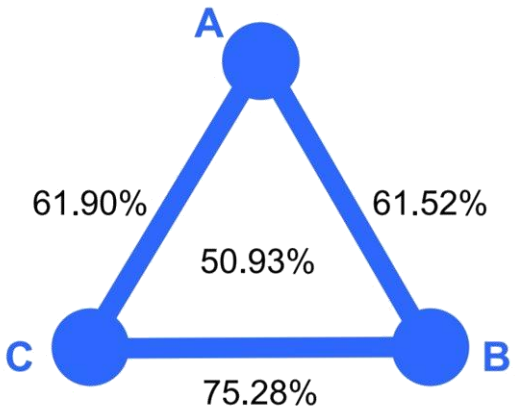
1. Predict label of tweets using pretrained model.
2. Give priority to tweets with highest model uncertainty.

Annotator agreement is low

Average pairwise agreement: **66.23%**

Average overall agreement: **50.93%**

Agreement of annotators with themselves: **90.37%**
(for identical tweets)



Bias in our ground truth set

- Labelling
  - Observed a linked with socio-demo's (gender, age, location)
  - People with similar socio-demo's tend to label tweets more similarly (and vice-versa)
  - --> Need for a labelling panel representative of the Flemish Twitter user? OR
  - --> Need to have someone similar as the author label the tweet?

Bias in our ground truth set

- Population bias
  - Twitter population is not representative of flemish population
  - Extra information or corrections are needed
- We can predict for a twitter profile
  - Gender with 92% accuracy
  - Province with 74% accuracy
  - Age category with 55% accuracy
  - --> starting point to do any kind of corrections/targeted labelling (but more work needed)

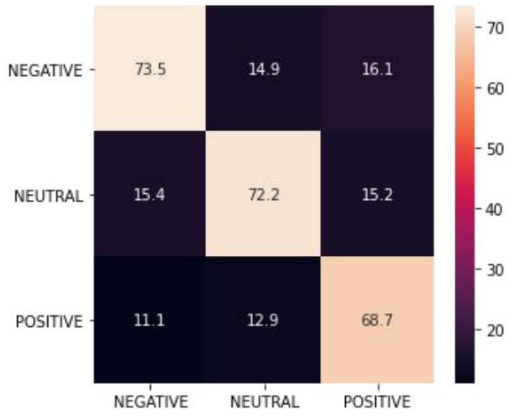
Model = RoBERTa transformer

High-level pipeline

Performance

71.5% average pairwise model agreement w.r.t. human annotators

(better than lexicon-based and 'traditional machine learning methods')

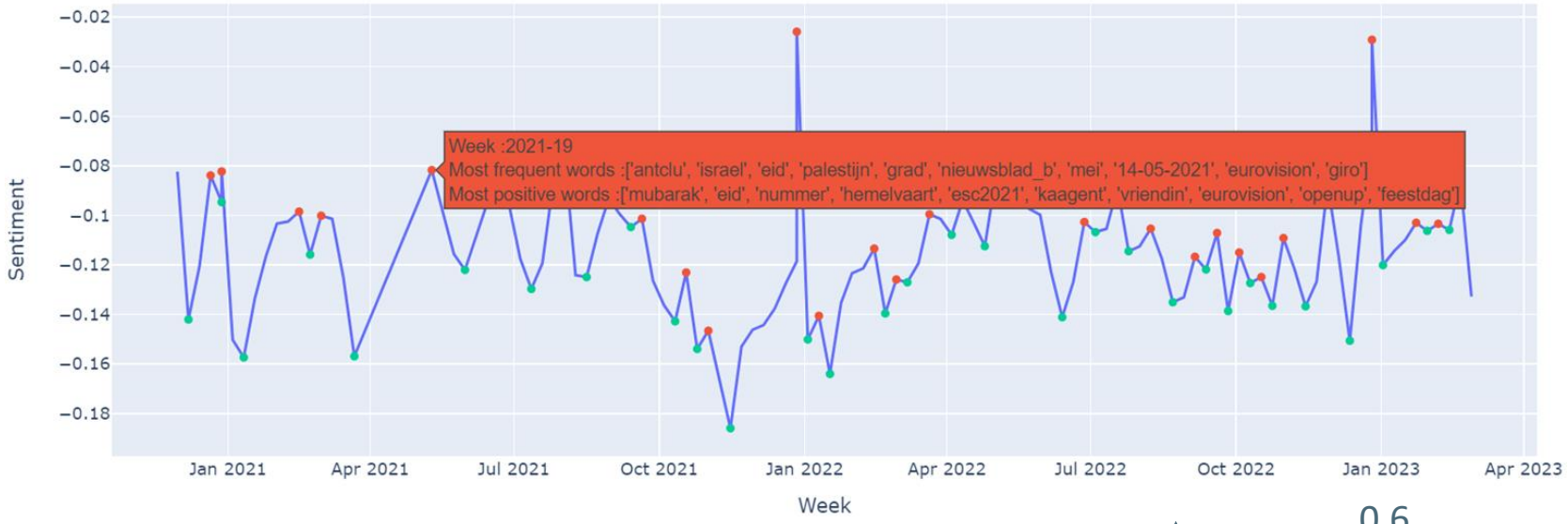


STATISTIEK VLAANDEREN

Vlaanderen is sterk in cijfers

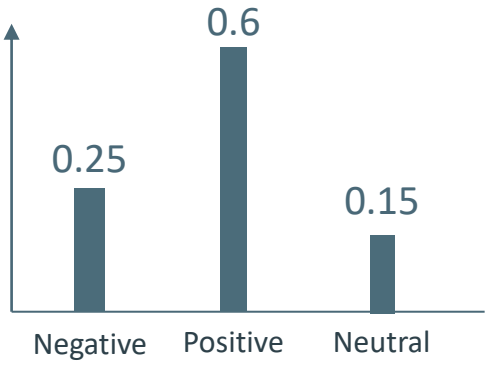
Insights: local peak detection

Weekly-level sentiment with positive and negative peaks:



How to select ‘most positive words’:

1. Create list of top N most frequent words in this week
2. Calculate average positivity of each word
3. Show top K most positive words



$$\text{positivity} = 0.6 - 0.25 = 0.35$$

STATISTIEK VLAANDEREN

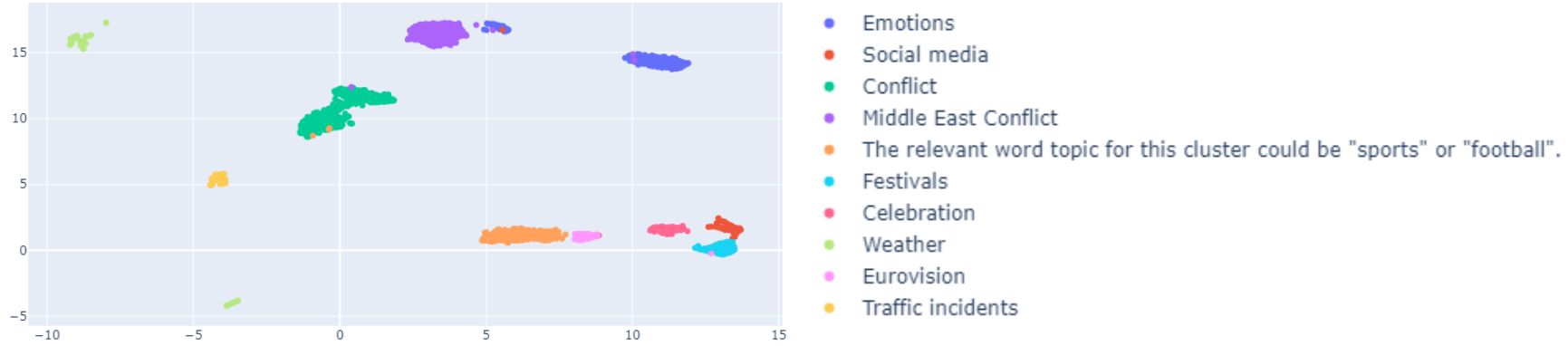
Vlaanderen is sterk in cijfers

Insights: topic / event clusters

Clustering algorithm, for each week

1. Project all tweet embeddings in a low-D space using UMAP
2. Cluster all tweets using HDBSCAN in this low-D space
3. Remove all tweets that do not belong to any cluster (“noise”)
4. For each cluster:
    1. Calculate top N most frequent words
    2. Request GPT to assign a single semantic label based on the top N words

2D projection of all tweets, colored by cluster



STATISTIEK VLAANDEREN

Vlaanderen is sterk in cijfers

Twitter (X) in the future?

- 2022: Elon Musk acquires Twitter
- May 2023: all existing API offerings are suspended. Free tiers are not useful, premium tiers are very costly and prone to changes
- June 2023: Linda Yaccarino becomes new Twitter CEO
- 3 July 2023: Twitter limits app usage to combat scrapers and crawlers
- 5 July 2023: Meta launches Threads;
- Our decision: discontinue the statistic (for now)
  - We cannot trust the availability and consistency of this dataset for the production official statistics

Lack of control is an insidious risk

with great amounts of external data
comes
great powerlessness!

C. De Boom & M. Reusens.
Changing Data Sources in the Age of Machine Learning for Official Statistics.
UNECE ML for Official Statistics Workshop, 2023.

Risk mitigation strategies
should be front and center
in your data science agenda and practices!

Should we use social media data for official stats?

- Quality dimensions (https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646)
  - Relevance
    - Yes, it contains information on important domains
  - Accuracy and Reliability
    - Accuracy: Requires more research on statistical properties
    - Reliability: Requires stronger agreements with data provider
  - Timeliness and punctuality
    - Yes, one of the key benefits
  - Comparability and coherence
    - Between regions: requires shared methods
    - Over time: requires more research
  - Accessibility and clarity
    - Accessibility: not different than other stats
    - Requires a better understanding of which real-life constructs are actually measured

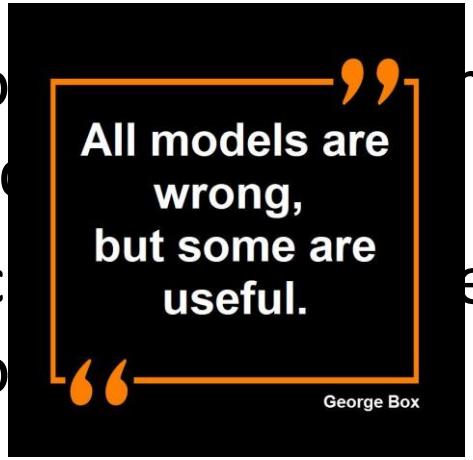
Should we use social media data for official stats?

- "YES!" ???
  - There are opportunities for value
  - Requires more research and joint work within the statistical community
- "NO!" ???
  - There are obvious properties making social media data imperfect datasets
  - Imperfections can be overcome, this is one of the core activities of statisticians

Should we use social media data for official stats?

- "YES!" ???
  - There are oppo ue
  - Requires more r nt work within the statistical comm

- "NO!" ???
  - There are obvio naking social media data imperfect
  - Imperfections c e, this is one of the core activities o



Key lessons

- New data sources/methods require lots of research before they are well understood
  - Collaborate with others, share the research load
  - Academia is a valuable partner here

- Risks related to external data sources
  - Identify and mitigate before starting any project

Questions?

- For all questions about the data science work at statistics Flanders
  - vsadatascience@vlaanderen.be