STATISTICS
FLANDERS

Statistiek Vlaanderen Data Science Hub & gebruik van twitter data voor officiële statistieken

Michael Reusens, Data Science Coördinator @ Statistics Flanders

SV-seminarie Donderdag 18 maart 2021

Flanders
State of the Art

Overzicht

- Data science hub
  - Doelstellingen
  - Activiteiten

- Vlaams Twitter sentiment
  - Opzet
  - Methode
  - Volgende stappen

Data science hub

- Doelstellingen
  1. Nieuwe/betere statistieken over Vlaanderen met data science technieken en nieuwe databronnen
  2. Data science opleidingsaanbod VO
  3. Netwerk

- Worden stapsgewijs uitgewerkt

Data science hub – Huidige samenwerkingen

- VO Overheidsdienst ←→ Hub

- Kennisinstellingen
  - Masterproeven en stage rond data science voor officiële statistieken
  - Officiële onderzoeksamenwerking met KU Leuven

- Internationaal
  - Andere statistiekbureaus
  - ONS-UNECE ML2021 group

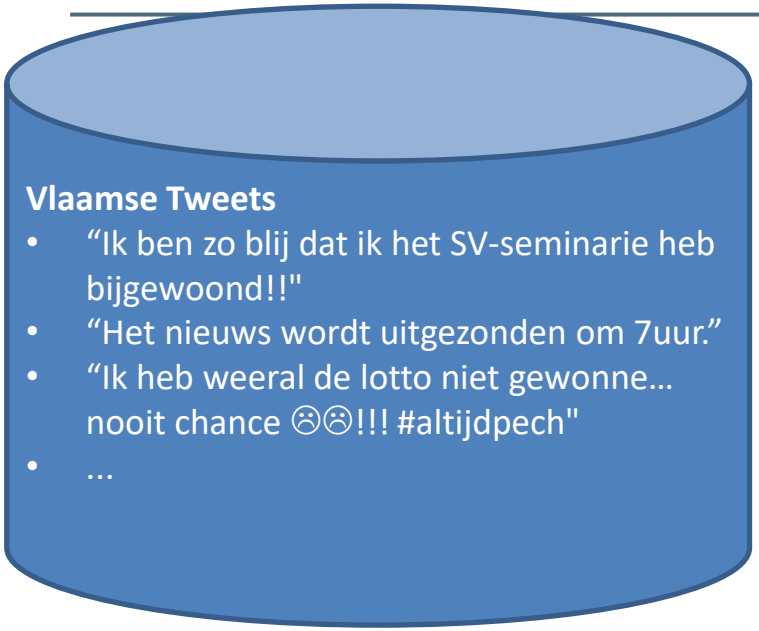
Data science hub – lopende experimenten

- Webscraping + Machine learning (ML) voor betere bedrijfsstatistieken
- Luchtfoto's + ML voor het bepalen van beroepsinkomen op sectorniveau
- Beter in kaart brengen van 'levenslang leren' met Google Trends en Twitter data
- Vlaams Twitter sentiment

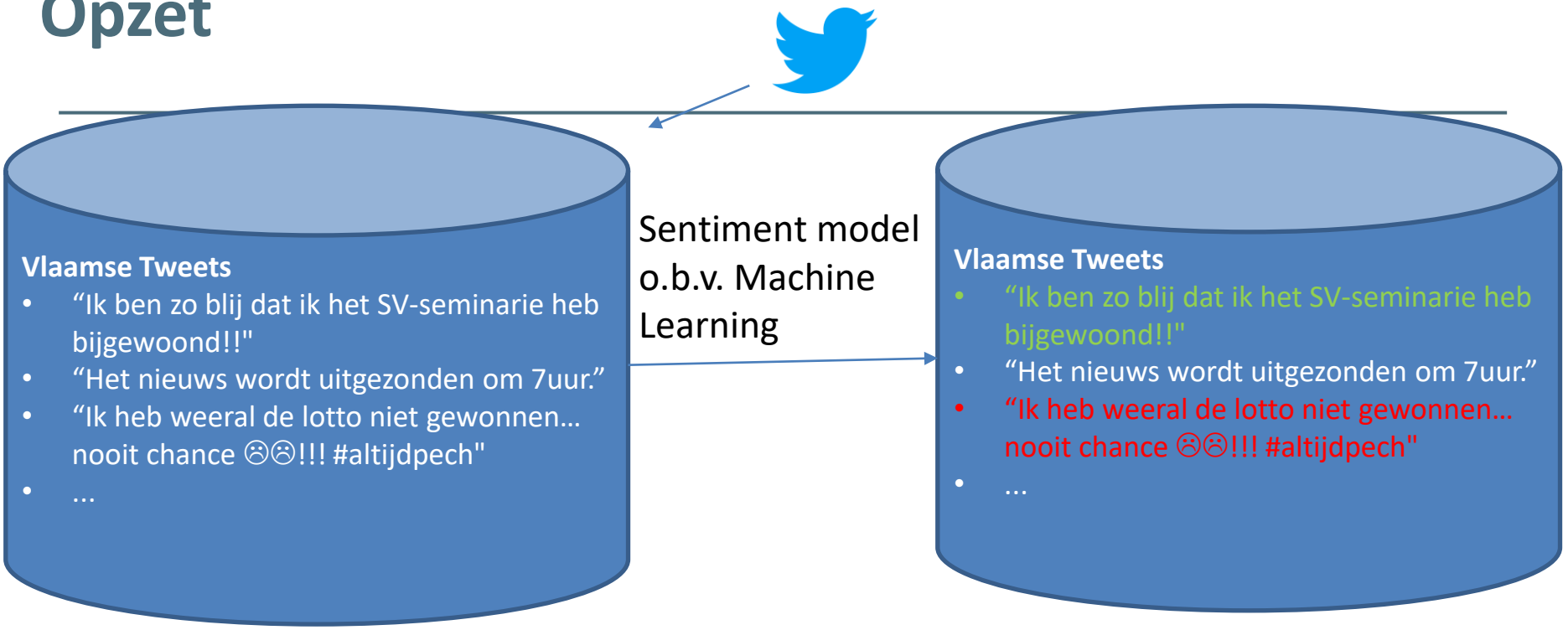
Algemeen Vlaams Twitter sentiment

- Begonnen als een alternatieve “perceived quality of life" statistiek
  - Hoge frequentie
  - Quasi real-time
- Eerste interne versie ontwikkeld
  - i.s.m. consultancy-partner
  - Als basis voor verdere verbeteringen en onderzoek

Opzet


Vlaamse Tweets
• “Ik ben zo blij dat ik het SV-seminarie heb bijgewoond!!"
• “Het nieuws wordt uitgezonden om 7uur.”
• “Ik heb weeral de lotto niet gewonne… nooit chance ☹☹!!! #altijdpech"
• …

Opzet


Vlaamse Tweets
• "Ik ben zo blij dat ik het SV-seminarie heb bijgewoond!!"
• "Het nieuws wordt uitgezonden om 7uur."
• "Ik heb weeral de lotto niet gewonnen… nooit chance ☹☹!!! #altijdpech"
• …
Sentiment model o.b.v. Machine Learning
Vlaamse Tweets
• "Ik ben zo blij dat ik het SV-seminarie heb bijgewoond!!"
• "Het nieuws wordt uitgezonden om 7uur."
• "Ik heb weeral de lotto niet gewonnen… nooit chance ☹☹!!! #altijdpech"
• …

Opzet

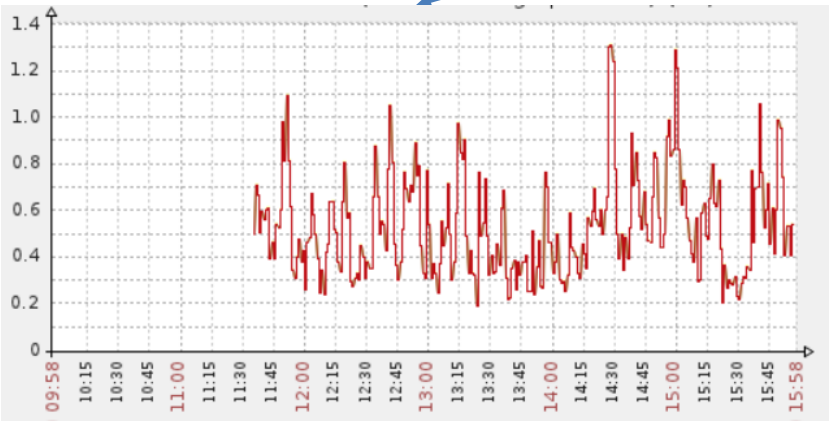
Vlaamse Tweets

- “Ik ben zo blij dat ik het SV-seminarie heb bijgewoond!!"
- “Het nieuws wordt uitgezonden om 7uur.”
- “Ik heb weeral de lotto niet gewonnen… nooit chance ☹☹!!! #altijdpech"
- …

Sentiment model o.b.v. Machine Learning

Vlaamse Tweets

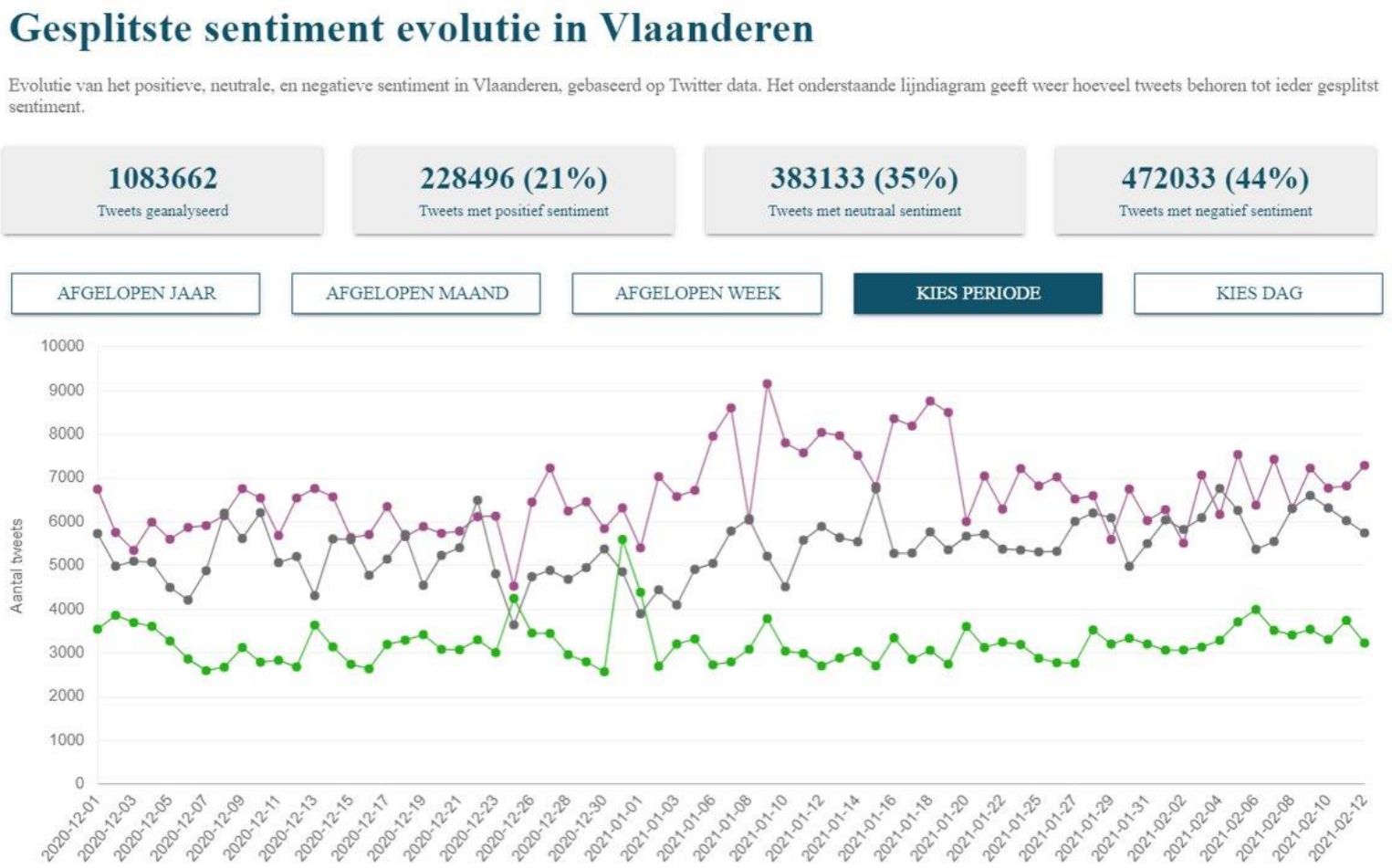
- “Ik ben zo blij dat ik het SV-seminarie heb bijgewoond!!"
- “Het nieuws wordt uitgezonden om 7uur.”
- “Ik heb weeral de lotto niet gewonnen… nooit chance ☹☹!!! #altijdpech"
- …



STATISTICS FLANDERS

Flanders State of the Art

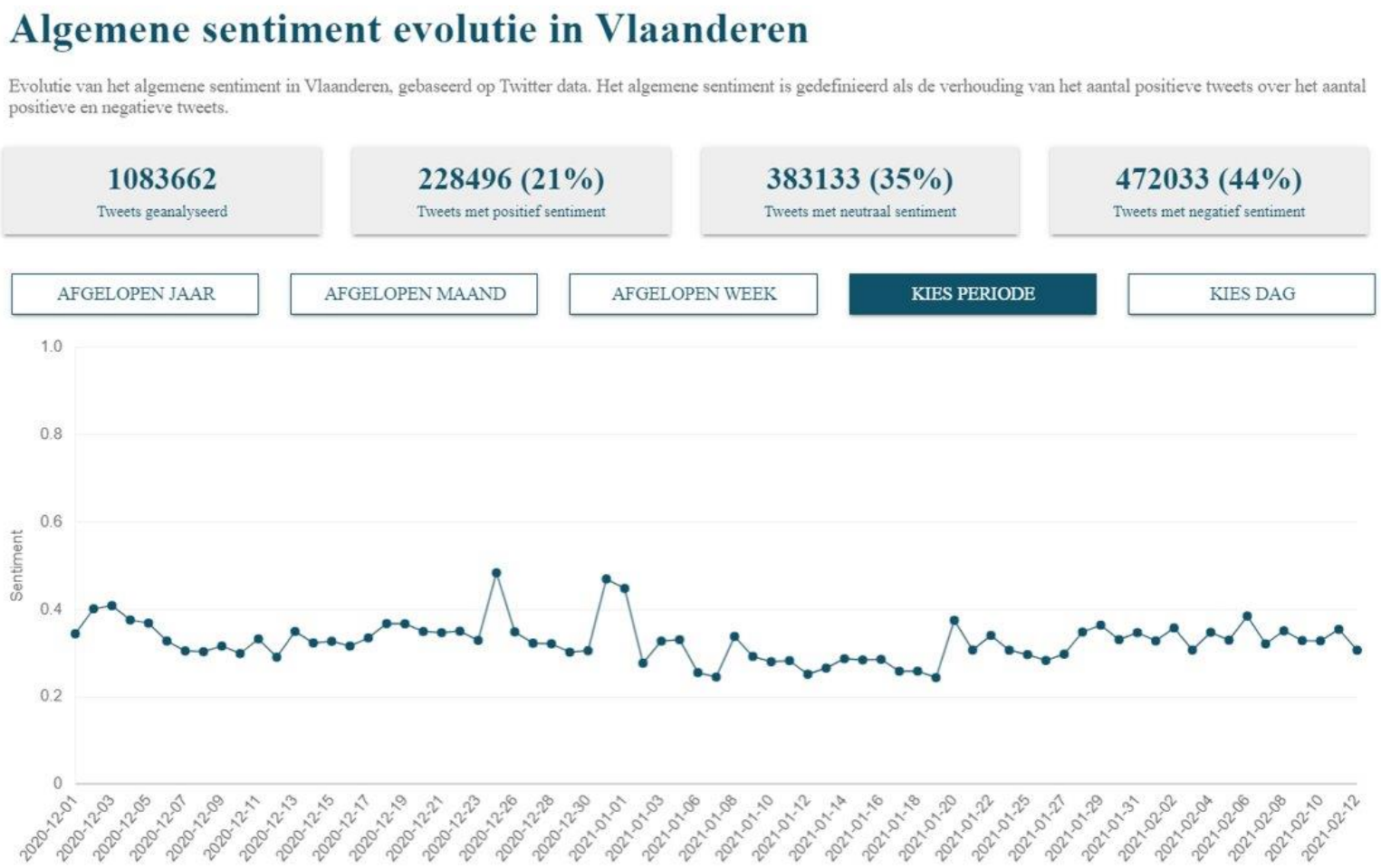
Demo – interne versie



STATISTICS
FLANDERS

Flanders
State of the Art

Demo



STATISTICS
FLANDERS

Flanders
State of the Art

De temperatuur van Vlaanderen

STATISTICS
FLANDERS

Flanders
State of the Art

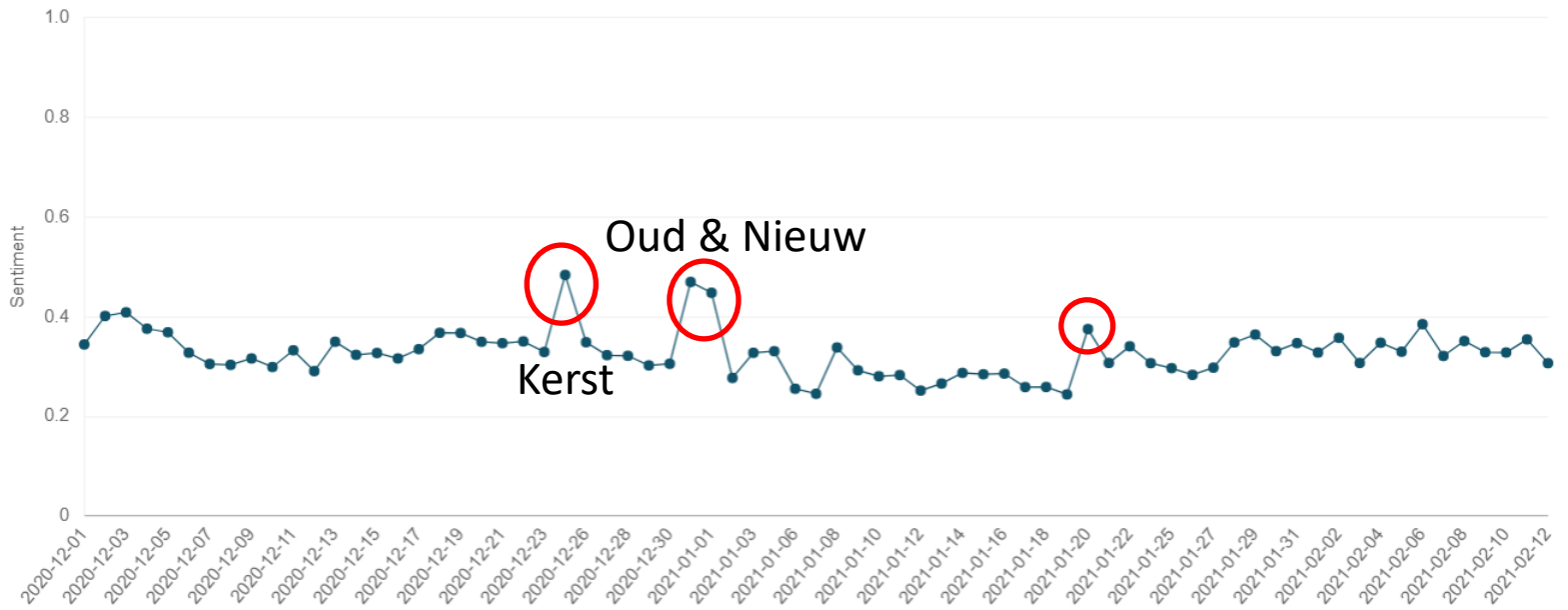
Demo

Algemene sentiment evolutie in Vlaanderen

Evolutie van het algemene sentiment in Vlaanderen, gebaseerd op Twitter data. Het algemene sentiment is gedefinieerd als de verhouding van het aantal positieve tweets over het aantal positieve en negatieve tweets.

1083662	228496 (21%)	383133 (35%)	472033 (44%)
Tweets geanalyseerd	Tweets met positief sentiment	Tweets met neutraal sentiment	Tweets met negatief sentiment

AFGELOPEN JAAR | AFGELOPEN MAAND | AFGELOPEN WEEK | KIES PERIODE | KIES DAG



STATISTICS FLANDERS

Flanders State of the Art

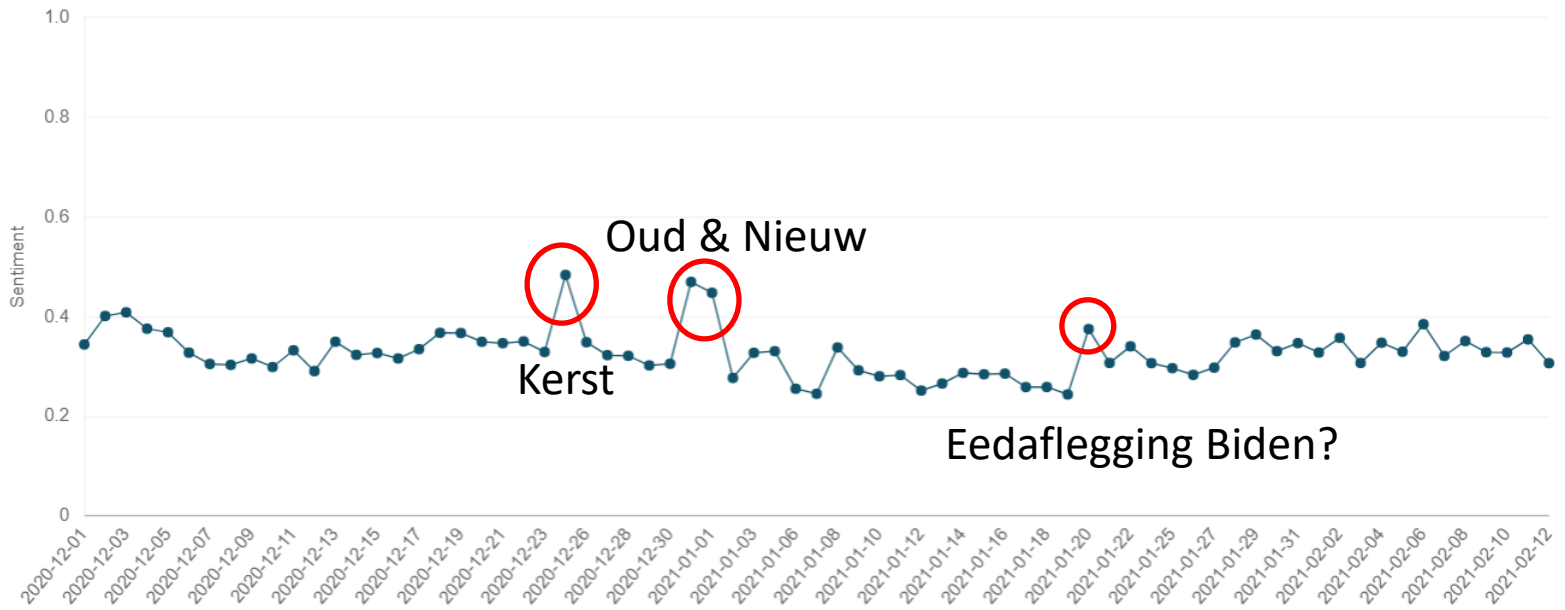
Demo

Algemene sentiment evolutie in Vlaanderen

Evolutie van het algemene sentiment in Vlaanderen, gebaseerd op Twitter data. Het algemene sentiment is gedefinieerd als de verhouding van het aantal positieve tweets over het aantal positieve en negatieve tweets.

1083662	228496 (21%)	383133 (35%)	472033 (44%)
Tweets geanalyseerd	Tweets met positief sentiment	Tweets met neutraal sentiment	Tweets met negatief sentiment

AFGELOPEN JAAR | AFGELOPEN MAAND | AFGELOPEN WEEK | KIES PERIODE | KIES DAG



STATISTICS FLANDERS

Flanders State of the Art

Demo

STATISTICS
FLANDERS

Flanders
State of the Art

Demo

STATISTICS
FLANDERS

Flanders
State of the Art

Data verzameling

- Twitter developer premium API
  - Quota op aantal tweets (8k/dag)
  - Opgesplitst naar 16 queries

- Benadering "Vlaamse Tweets"
  - Belgische accounts
  - Nederlandstalige tweet

Sentiment annotatie – training data

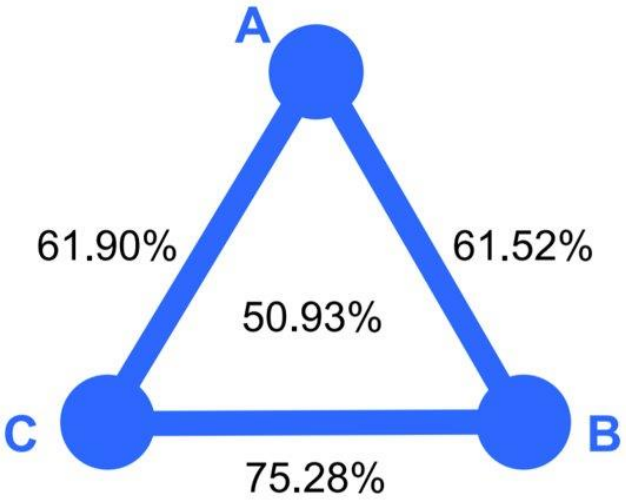
- Gelabelde dataset opgesteld
  - 3 job-studenten
  - Tweets van de afgelopen 5 jaar
- Intelligente annotatie strategie
  - Active learning
  - Diversiteit topics
- 50.000 geannoteerde Vlaamse Tweets

Machine learning model

- Eigen sentiment-model getraind
  - Meer controle en studie-potentieel
- RobBERT Transformer model
  - https://dtai.cs.kuleuven.be/stories/post/pieter-delobelle/robbert/
- 71.4% "accuracy"
  - = Overeenkomst met job studenten

Overeenkomst met annotator

- Nood aan uitgebreide annotator methode
  - Paarsgewijze overeenkomst : 66.2% gemiddeld
  - Unanieme overeenkomst: 50.9% gemiddeld



  - Overeenkomst met zichzelf: 90.4% gemiddeld

Verder onderzoek

- Samenwerking met GNH-project en kennisinstituten
- Selectiviteit van Twitter gebruikers
- Sentiment annotatie methode
- Machine learning model
  - Tevreden met architectuur
  - Wensen beter begrip over gedrag
- Toepassingen van deze statistiek

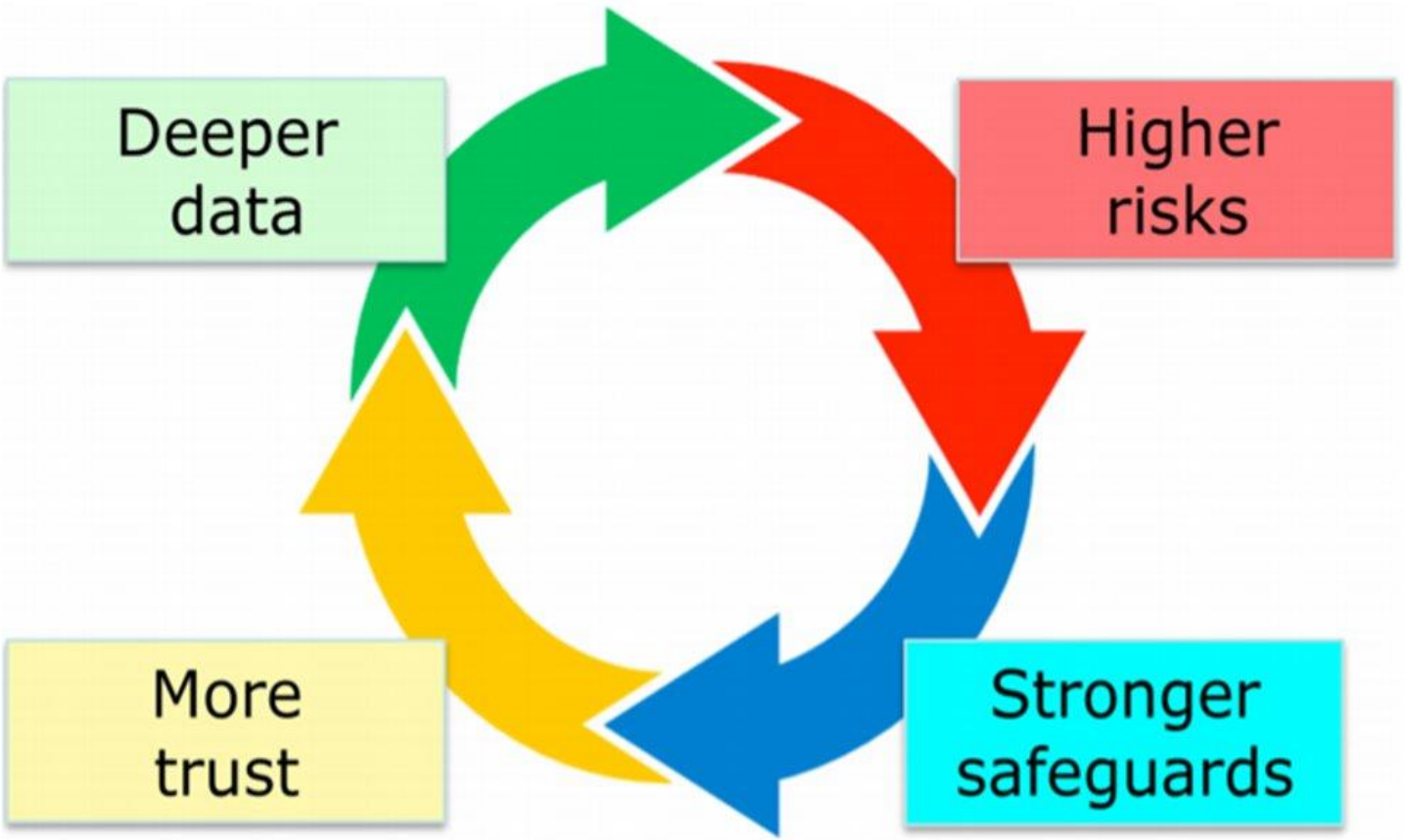
Toepassingen

- Sentiment over specifieke topics
  - Levenslang leren in Vlaanderen
  - →
  - …

STATISTICS FLANDERS

Flanders State of the Art

The smart and trusted cycle



Source: Ricciato e.a., 2019

Bedankt voor uw aandacht!

- Vragen – suggesties – ideeën voor/nood aan nieuwe en betere Vlaamse statistieken?

- Contact: michael.reusens@vlaanderen.be