

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

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 onderzoekssamenwerking 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

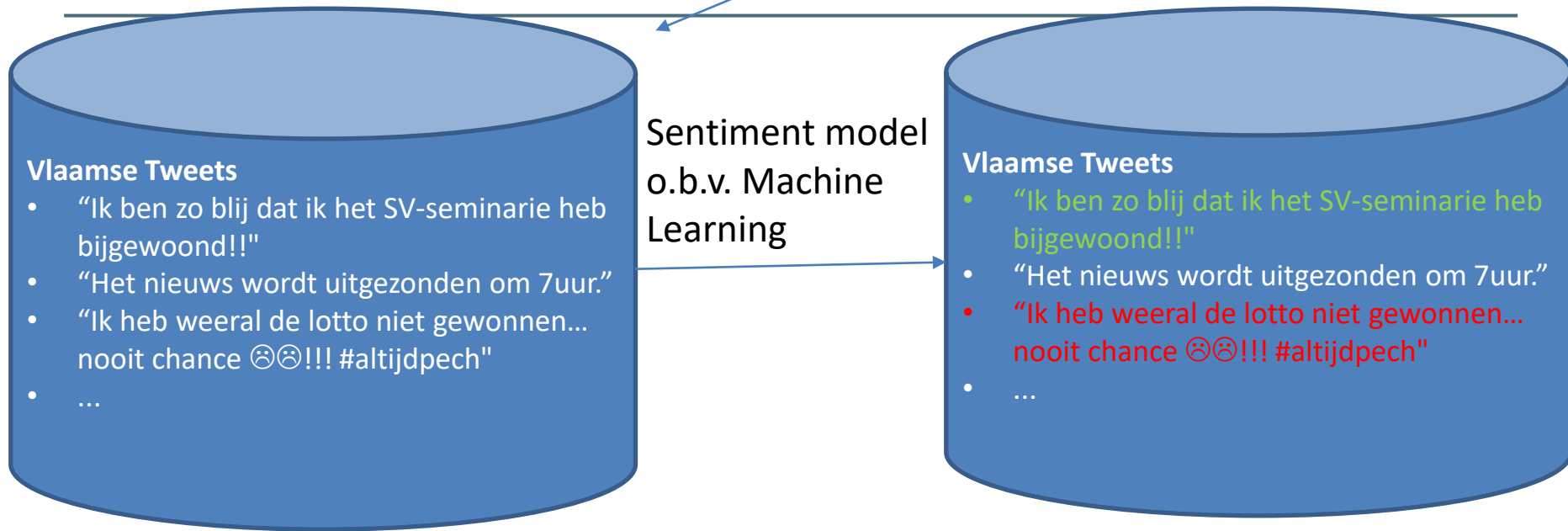
Opzet



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



Opzet



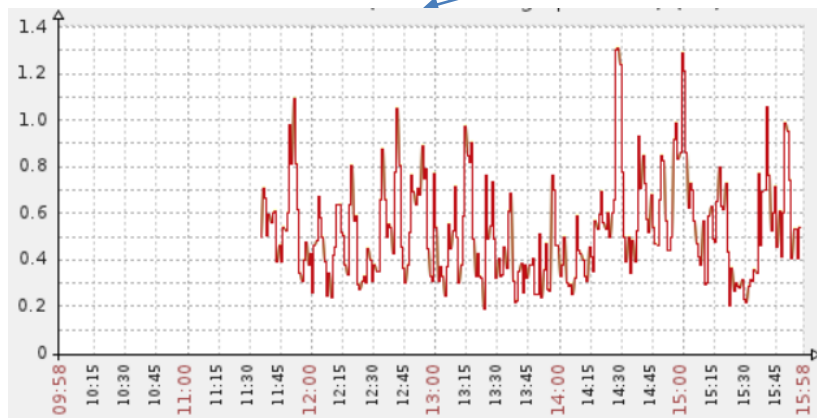
Vlaamse Tweets

- "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"
- ...



Demo – interne versie

Gesplitste sentiment evolutie in Vlaanderen

Evolutie van het positieve, neutrale, en negatieve sentiment in Vlaanderen, gebaseerd op Twitter data. Het onderstaande lijndiagram geeft weer hoeveel tweets behoren tot ieder gesplitst sentiment.

1083662

Tweets geanalyseerd

228496 (21%)

Tweets met positief sentiment

383133 (35%)

Tweets met neutraal sentiment

472033 (44%)

Tweets met negatief sentiment

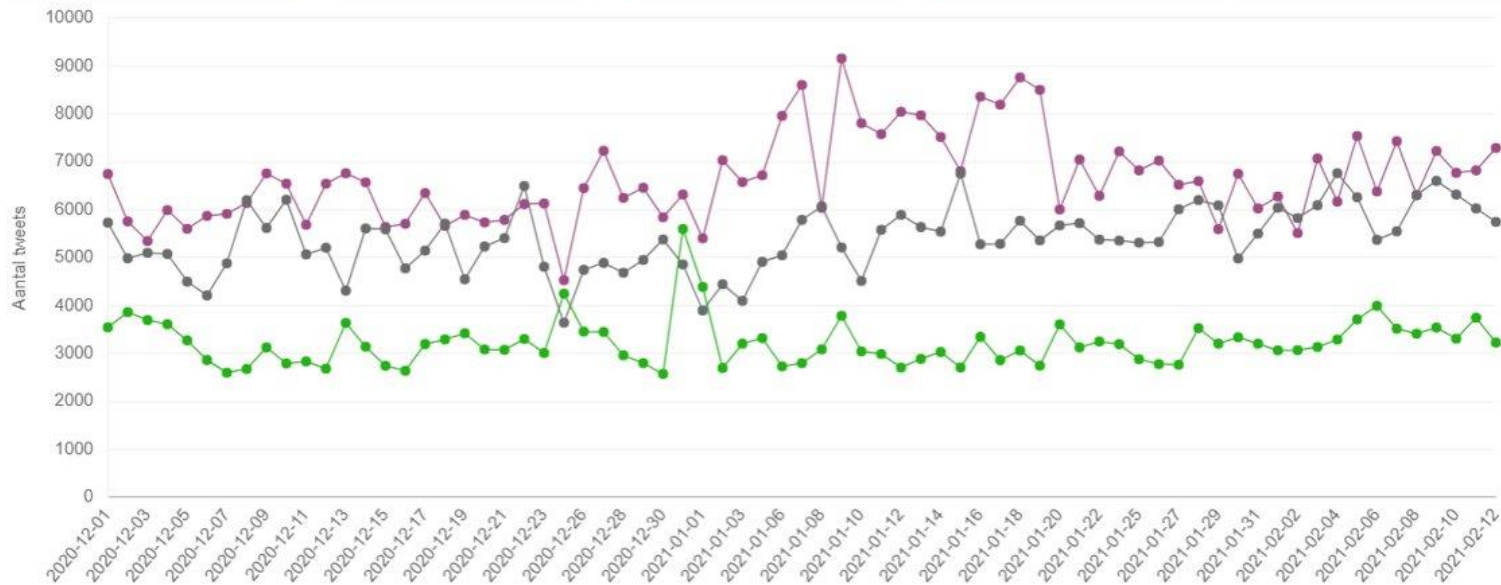
AFGELOPEN JAAR

AFGELOPEN MAAND

AFGELOPEN WEEK

KIES PERIODE

KIES DAG



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

Tweets geanalyseerd

228496 (21%)

Tweets met positief sentiment

383133 (35%)

Tweets met neutraal sentiment

472033 (44%)

Tweets met negatief sentiment

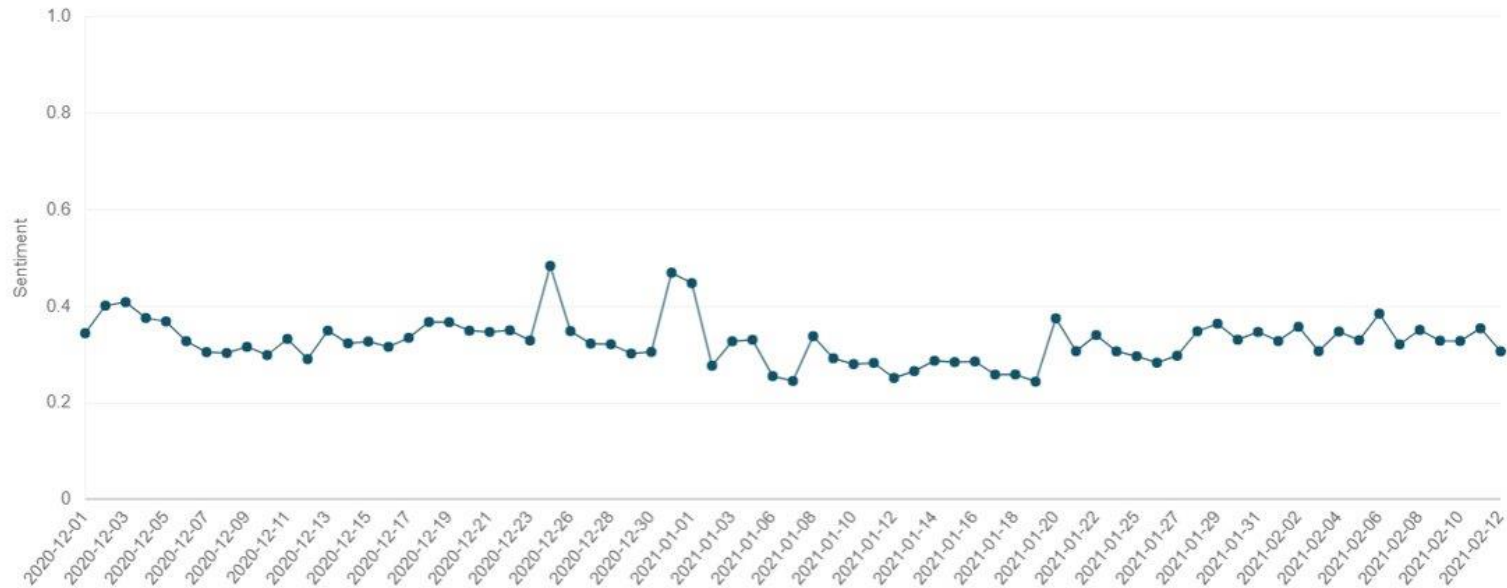
AFGELOPEN JAAR

AFGELOPEN MAAND

AFGELOPEN WEEK

KIES PERIODE

KIES DAG



De temperatuur van Vlaanderen



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

Tweets geanalyseerd

228496 (21%)

Tweets met positief sentiment

383133 (35%)

Tweets met neutraal sentiment

472033 (44%)

Tweets met negatief sentiment

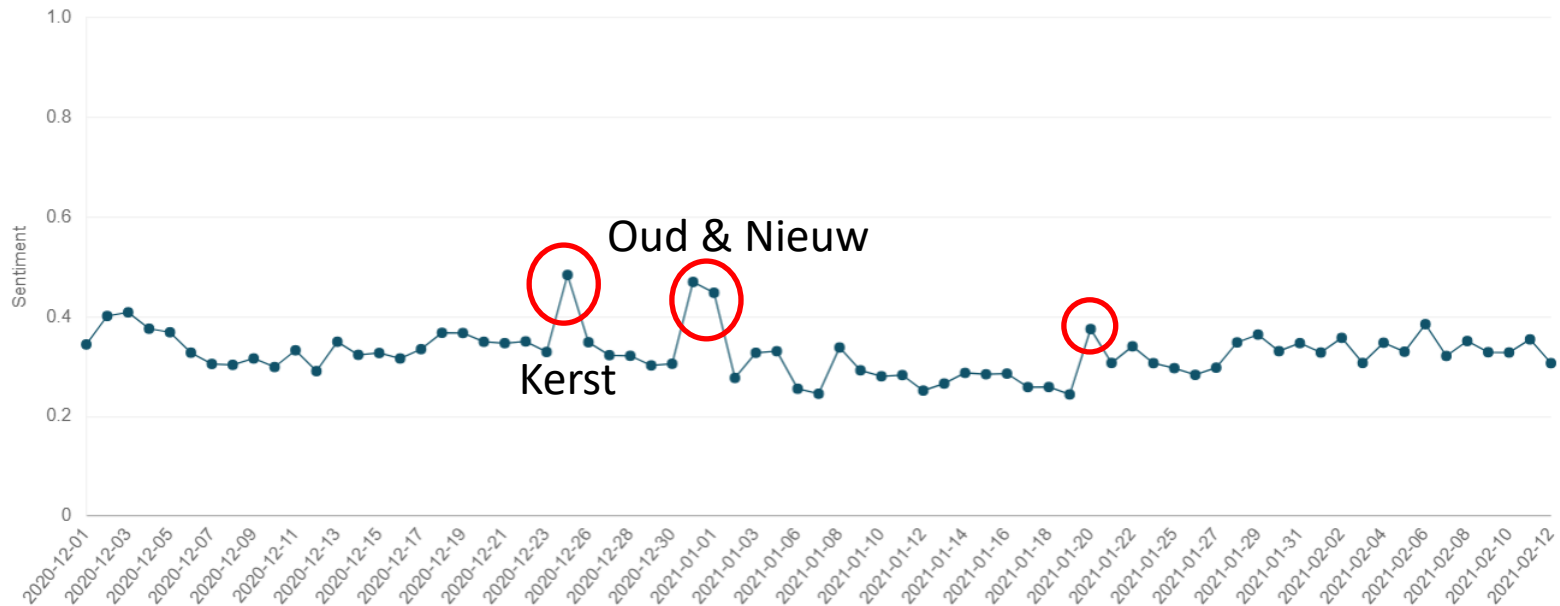
AFGELOPEN JAAR

AFGELOPEN MAAND

AFGELOPEN WEEK

KIES PERIODE

KIES DAG



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

Tweets geanalyseerd

228496 (21%)

Tweets met positief sentiment

383133 (35%)

Tweets met neutraal sentiment

472033 (44%)

Tweets met negatief sentiment

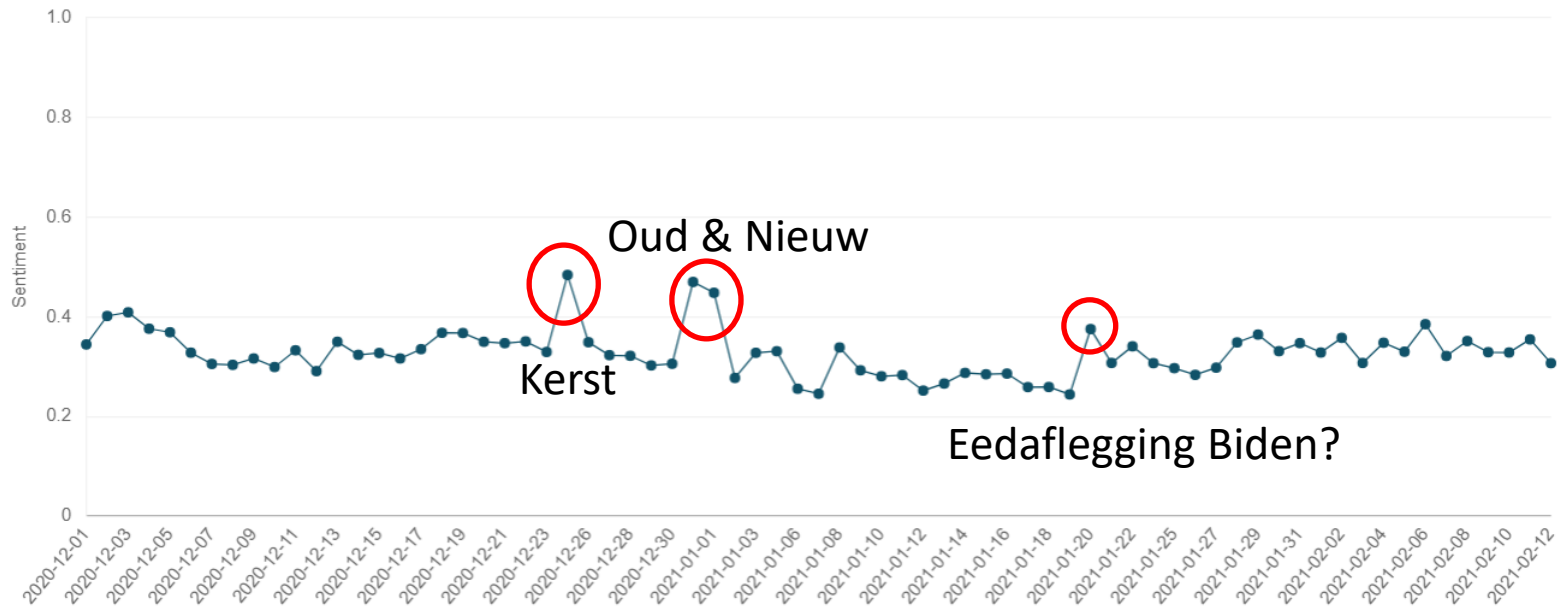
AFGELOPEN JAAR

AFGELOPEN MAAND

AFGELOPEN WEEK

KIES PERIODE

KIES DAG



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.

15283

Tweets geanalyseerd

3607 (24%)

Tweets met positief sentiment

5671 (37%)

Tweets met neutraal sentiment

6005 (39%)

Tweets met negatief sentiment

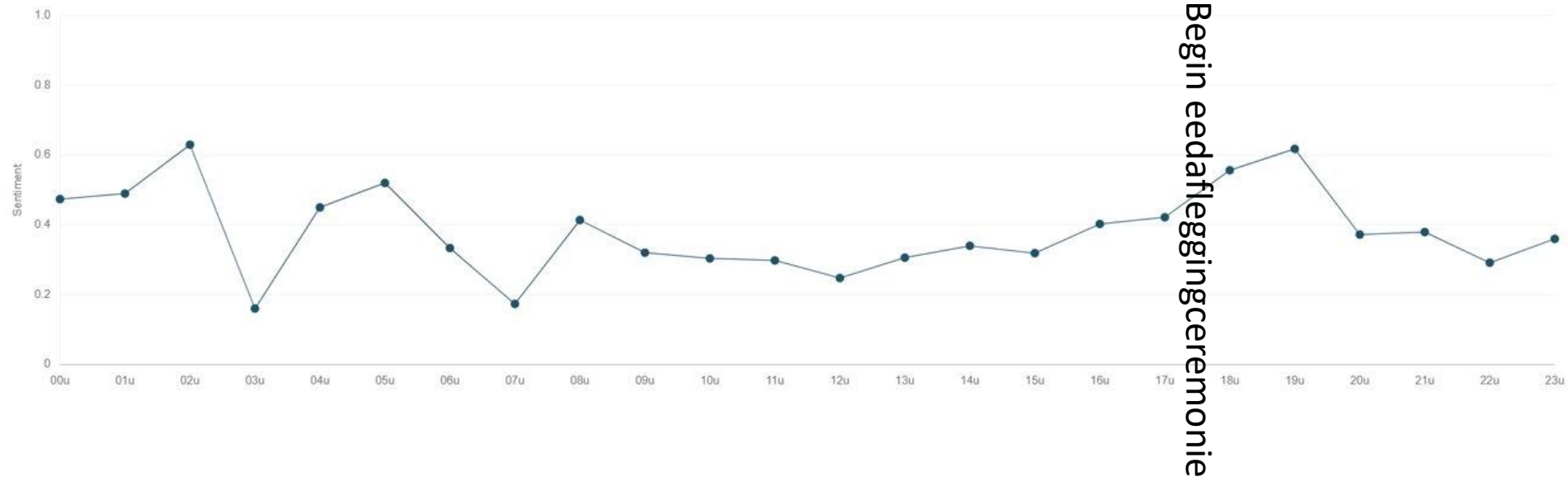
AFGELOPEN JAAR

AFGELOPEN MAAND

AFGELOPEN WEEK

KIES PERIODE

KIES DAG



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.

1352290

Tweets geanalyseerd

285832 (21%)

Tweets met positief sentiment

483740 (36%)

Tweets met neutraal sentiment

582718 (43%)

Tweets met negatief sentiment

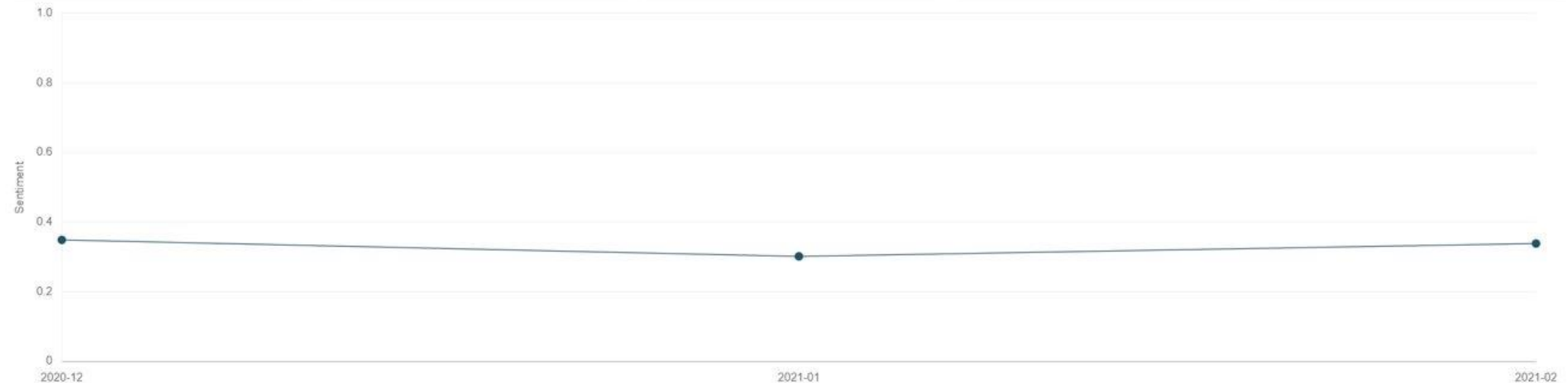
AFGELOPEN JAAR

AFGELOPEN MAAND

AFGELOPEN WEEK

KIES PERIODE

KIES DAG



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

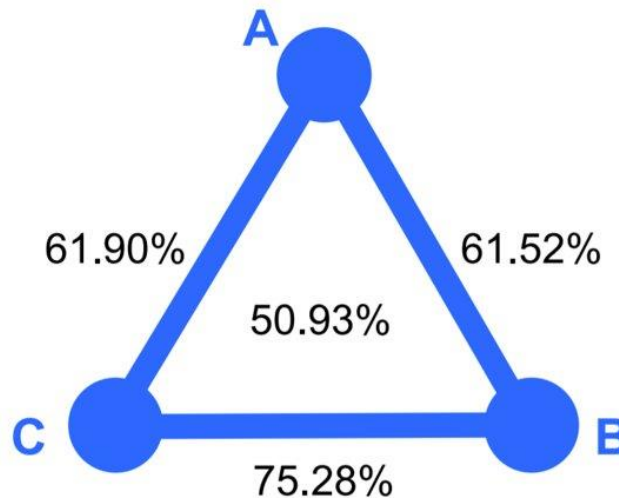
- Gelabelde dataset opgesteld
 - 3 job-studenten
 - Tweets van de afgelopen 5 jaar
- Intelligente annotatie strategie
 - Active learning
 - Diversiteit topics
- 50.000 geannoteerde Vlaamse Tweets

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

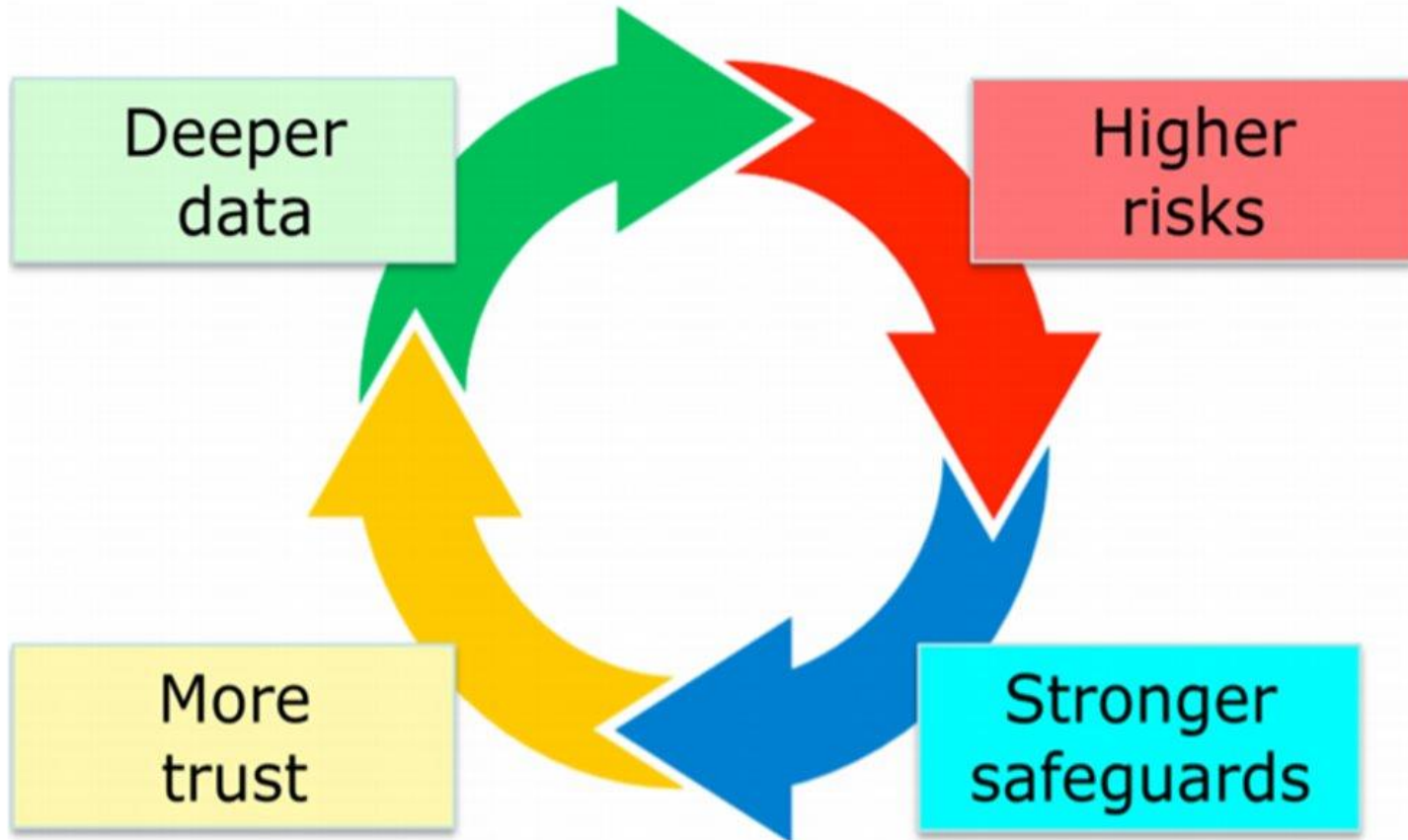
- Sentiment over specifieke topics
 - Levenslang leren in Vlaanderen



The screenshot shows a Reuters news article. At the top left is the Reuters logo. To the right are navigation links: World, Business, Markets, Breakingviews, Video, and more. Below the navigation is a horizontal line. Underneath, the article category 'TECHNOLOGY NEWS' is shown, followed by the date and time 'FEBRUARY 15, 2021 / 11:17 AM / UPDATED 22 DAYS AGO'. The main headline reads 'Twitter a goldmine for tracking consumer mood on prices, Bank of Italy finds'. Below the headline, it says 'By Stefano Bernabei' and '4 MIN READ'. There are social media icons for Facebook and Twitter. At the bottom of the article preview, the text reads: 'ROME (Reuters) - The Bank of Italy said on Monday a set of experimental indicators it created from the content of millions of tweets accurately tracked consumer mood on price, offering scope for a powerful new monetary policy tool.'



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