KU LEUVEN

Onderzoekssamenwerking tussen VSA en KU Leuven

Manon Reusens

Wie ben ik?

- Manon Reusens
  - 3de jaars PhD Student aan de KU Leuven
  - Promotor Bart Baesens
    Co-Promotoren: Seppe vanden Broucke, Wouter Verbeke
  - Faculteit Economie & Bedrijfswetenschappen
  - Leerstoel VSA
- Focus Onderzoek
  - Toepassingen van Natural Language Processing
    - Tekst classificatie
    - Sentiment Analyse: Twitter
    - Webscraping
    - Trefwoord extractie en verrijking
    - Bias in Large Language Models

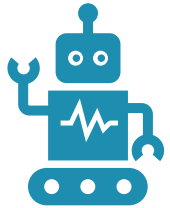
https://github.com/manon-reusens/

Manon Reusens

Overzicht Onderzoek

Twitter / X

Webscraping



Other

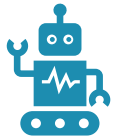
Overzicht Onderzoek

Twitter / X

- Predicting the Demographics of Twitter Users with Programmatic Weak Supervision
- Predicting Annotation Difficulty to Mitigate Annotator Bias
- Comparison of Different Modeling Techniques for Flemish Twitter Sentiment Analysis
- Quality of Life in Flanders: A Comparative Study Using Twitter and Survey Data
- Measuring Political Confidence Using Twitter Sentiment Analysis: a Belgian example

Webscraping

- Web scraping company information for official statistics: Automatically finding company URLs and generating NACE codes
- Detecting industry trends using web scraping on Belgian company websites
- Topic Modeling applied on innovation studies of Flemish companies
- Automated URL retrieval for Analyses of Companies using Web Scraping & Machine Learning
- Recommending NACEBEL codes for companies based on their website text



Other

- Natural language interfaces to numerical and tabular data: querying for statistics
- Strava data in the context of local knowledge-driven sports and physical activity policy

Onderzoek naar: Twitter / X

Doel

- Kunnen Tweets gebruikt worden als bron voor het publiceren van officiële statistieken? Bv: is het een goede substituut of complement voor surveys?

Survey data:

- Kostelijk
- Tijdsrovend
- Dalende response rate
- Onderhevig aan bias
- Moeilijk om mensen te vinden voor zeldzame events

Social media data:

- Niet-intrusieve psychologische beoordelingen
- Verandering meten in de publieke opinie
- Mapt in real-time het sentiment van een natie
- Elimineert sociale wenselijkheidsbias

- Hoe krijg je een representatieve sample voor de volledige populatie o.b.v Twitter?
  - →Daarvoor moeten we de demografische gegevens hebben van Twitteraars om een representatieve sample te kunnen samenstellen

Predicting the Demographics of Twitter Users with Programmatic Weak Supervision

Jonathan Tonglet, Astrid Jehoul, Manon Reusens, Michael Reusens, and Bart Baesens

Tonglet, J., Jehoul, A., Reusens, M., Reusens, M., & Baesens, B. (2024). Predicting the demographics of Twitter users with programmatic weak supervision. TOP, 1-37.

Hoofdpunten

- Demografische labels verzamelen van Twitter gebruikers in Vlaanderen
  - Handmatig labelen is duur en niet schaalbaar
  - Er zijn geen grote datasets beschikbaar, zeker niet voor niet-Engelse talen
- Doel: accuraat voorspellen van demografische gegevens van Twitter gebruikers met minimale tussenkomst van labels
- Oplossing: **Three-step Programmatic Weak Supervision** om gender, leeftijd en locatie te voorspellen voor Vlaanderen

Doel

- Kunnen Tweets gebruikt worden als bron voor het publiceren van officiële statistieken? Bv: is het een goede substituut of complement voor surveys?

Survey data:

- Kostelijk
- Tijdsrovend
- Dalende response rate
- Onderhevig aan bias
- Moeilijk om mensen te vinden voor zeldzame events

Social media data:

- Niet-intrusieve psychologische beoordelingen
- Verandering meten in de publieke opinie
- Mapt in real-time het sentiment van een natie
- Elimineert sociale wenselijkheidsbias

- Hoe bereken je het Twitter sentiment?
  - Welke modellen kunnen hiervoor gebruikt worden?

Comparison of Different Modeling Techniques for Flemish Twitter Sentiment Analysis

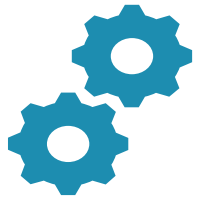
Manon Reusens, Michael Reusens, Marc Callens, Seppe vanden Broucke, and Bart Baesens

Reusens, M., Reusens, M., Callens, M., vanden Broucke, S., & Baesens, B. (2022). Comparison of Different Modeling Techniques for Flemish Twitter Sentiment Analysis. Analytics, 1(2), 117-134.

Hoofpunten

- Case study voor Vlaamse sentiment analyse
- Vergelijken van 4 verschillende categorieën van modellen
  - Lexicon-based methodes
  - Traditionele Machine Learning
  - Neurale Netwerken
  - Attention-based methoden
- De meest geavanceerde methoden tonen betere resultaten dan simpelere methoden.

Lexicon-Based Methods



Traditional Machine Learning Models

Neural Networks

Attention-Based Models

Doel

- Kunnen Tweets gebruikt worden als bron voor het publiceren van officiële statistieken? Bv: is het een goede substituut of complement voor surveys?

Survey data:

- Kostelijk
- Tijdsrovend
- Dalende response rate
- Onderhevig aan bias
- Moeilijk om mensen te vinden voor zeldzame events

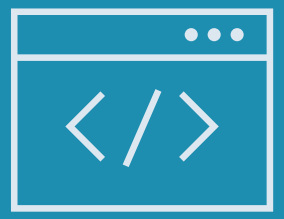
Social media data:

- Niet-intrusieve psychologische beoordelingen
- Verandering meten in de publieke opinie
- Mapt in real-time het sentiment van een natie
- Elimineert sociale wenselijkheidsbias

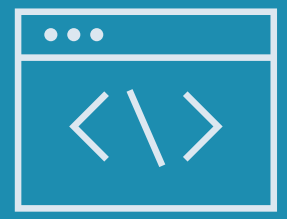
- Twee case studies waarbij Twitter sentiment data en survey data naast elkaar gelegd worden.

Hoofpunten

- Quality of life in Vlaanderen & vertrouwen in verschillende overheidsinstellingen
  - Twitter sentiment analyse naast survey resultaten gelegd.
- Twitter sentiment analyse kan gebruikt worden als complement voor survey data, niet als substituut
  - Wel correlatie



Onderzoek naar: Webscraping



Ultimate Goal

Focus op mensen

Focus op bedrijven

Kan website data gebruikt worden voor officiële statistieken?

Toevoegen van NACEBEL codes adhv website data

Automatisch industrie trends ontdekken adhv website data

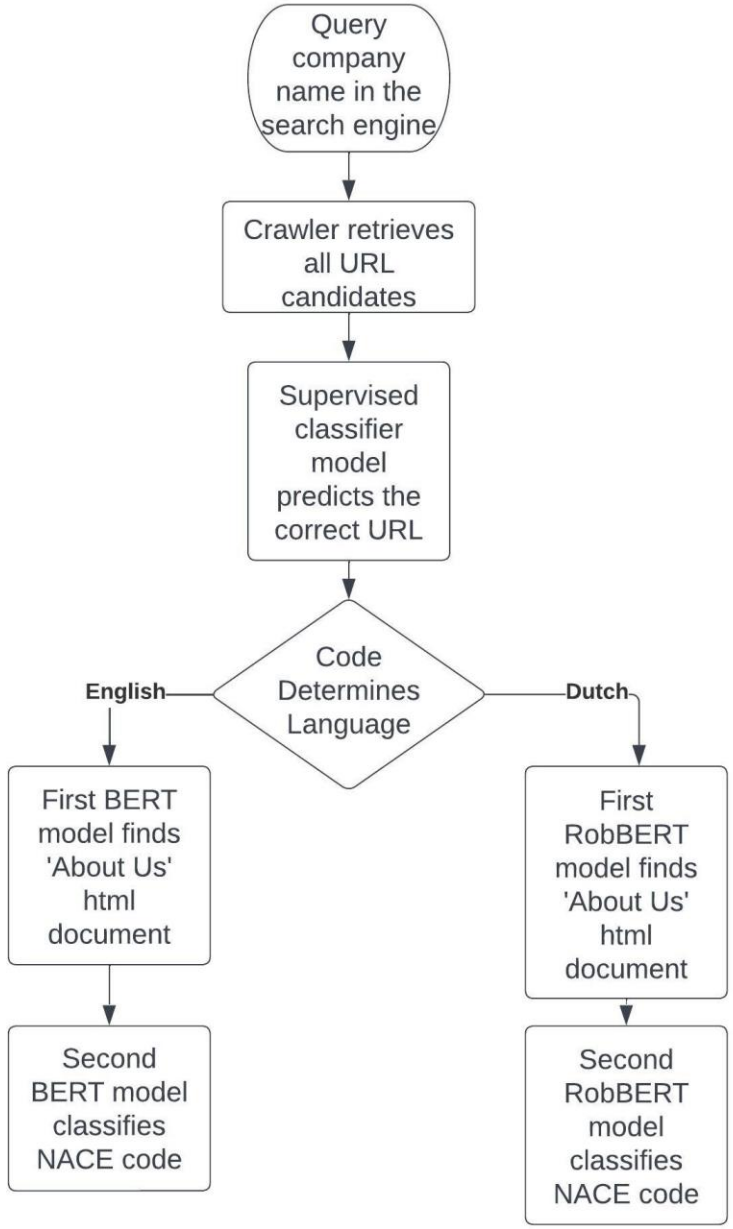
Maar assumptie is dat de **URLs** van de bedrijven **gekend** zijn

Web scraping company information for official statistics: Automatically finding company URLs and generating NACE codes

Luis Miguel Lara Zenozain, Ipek Senay, Hazem Ahmed, Manon Reusens, Michael Reusens, and Bart Baesens

Pipeline

- correcte URL vinden uit the kandidaat URLs
- ‘About us’ pagina’s & NACEBEL codes zijn verkregen via BERT-based modellen



Hoofdpunten

Vorig jaar: Werkende pipeline om URLs te vinden en bedrijfswebsites in NACEBEL codes te classificeren

Dit jaar: meer in depth onderzoek naar deze 2 facetten door ze op te splitsen

- Automated URL retrieval for Analyses of Companies using Web Scraping & Machine Learning
- Recommending NACEBEL codes for companies based on their website text

Ultimate Goal

Focus op mensen

Kan website data gebruikt worden voor officiële statistieken?

Toevoegen van NACEBEL codes adhv website data

Automatisch industrie trends ontdekken adhv website data

Twee studies:
Detecting industry trends on Belgian company websites
Topic modeling applied on innovation studies of Flemish companies

Topic modelling applied on innovation studies of Flemish companies

Annelien Crijns, Victor Vanhullebusch, Manon Reusens, Michael Reusens, and Bart Baesens

Crijns, A., Vanhullebusch, V., Reusens, M., Reusens, M., & Baesens, B. (2023). Topic modeling applied on innovation studies of Flemish companies. *Journal of Business Analytics,* 1-12.

Hoofdpunten

Analyseren van innovatieve en non-innovatieve bedrijven adhv hun website data – Top2Vec

Onder de innovatieve bedrijven werden de volgende sectoren opgemerkt: Engineering, IT and consultancy, Transport and Logistics, Food, Wood & furniture industry

Woorden gerelateerd aan innovatieve bedrijven: Technologie, leven, onderzoek, nieuws, oplossingen, sustainability, investor relations

Angelov, D. (2020). Top2vec: Distributed representations of topics. http://arxiv.org/abs/2008.09470
Schopf, T., Braun, D., & Matthes, F. (2021). Lbl2vec: An embedding-based approach for unsupervised document retrieval on predefined topics. In Proceedings of the 17th International Conference on web information systems and technologies - webist, Valletta, Malta, (p. 124–132). SciTePress.

Samenwerking

- Studenten HIR(B)/information management
  - Positieve feedback
  - Applicatiegericht → zeer enthousiast
  - Nut van hun thesis

Publicaties

- Crijns, A., Vanhullebusch, V., Reusens, M., Reusens, M., & Baesens, B. (2023). Topic modelling applied on innovation studies of Flemish companies. *Journal of Business Analytics*, 1-12.
- Reusens, M., Reusens, M., Callens, M., vanden Broucke, S., & Baesens, B. (2022). Comparison of Different Modeling Techniques for Flemish Twitter Sentiment Analysis. *Analytics*, *1*(2), 117-134.
- Tonglet, J., Jehoul, A., Reusens, M., Reusens, M., & Baesens, B. (2024). Predicting the Demographics of Twitter Users with Programmatic Weak Supervision. *TOP*, 1-37

Questions?