STATISTIEK VLAANDEREN

Sentiment analysis of Flemish tweets

Michael Reusens, Data Scientist @ Statistics Flanders
Marc Callens, Data Strategist @ Statistics Flanders

Vlaanderen
is sterk in cijfers

Overview

- Project goal
- Background
- Method overview
- Data collection
- Data processing
- Sentiment classification model
- Future work
- Conclusion

Project goal

- Binary sentiment classification of Flemish tweets

- Purpose
  - Creation of new quality of life statistic (complement to subjective wellbeing from survey?)
  - Get experience with Twitter data and ML

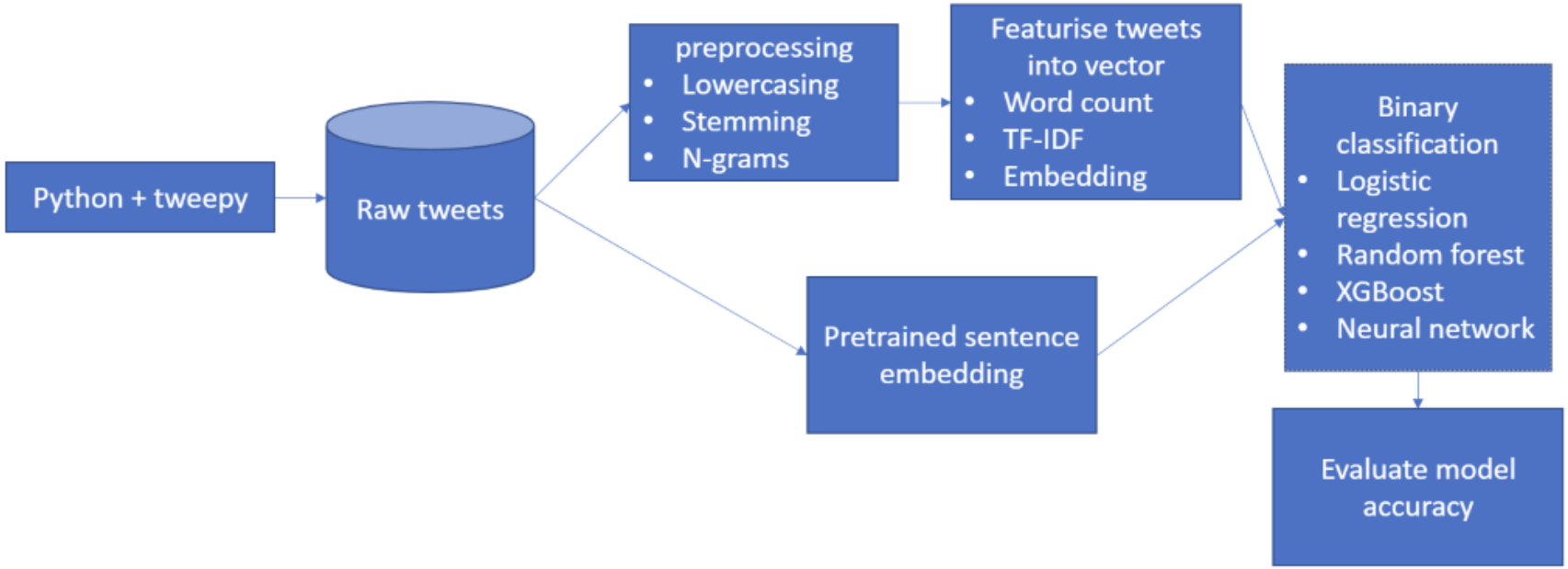
Background

- Statistics Flanders is a relatively small, young organisation
  - Created in 2016

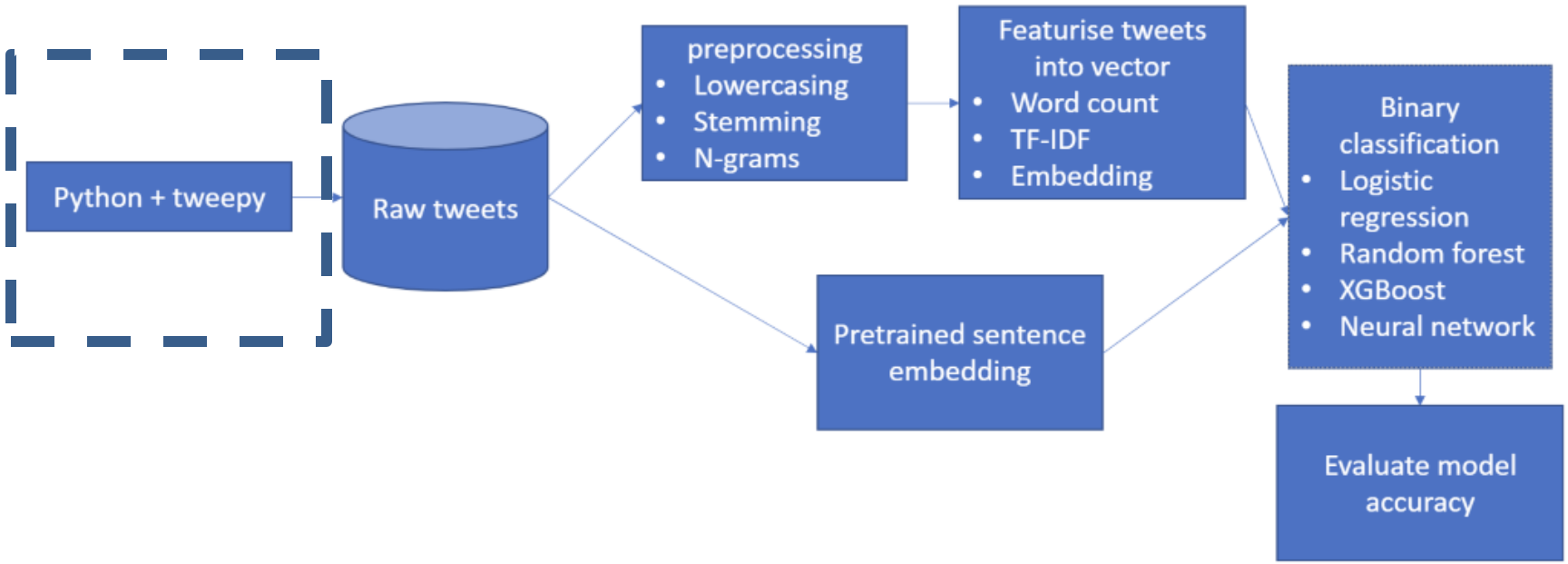
- Data innovation as one of our strategic pillars

- First experiments in using ML for better/new official statistics
  - Hired a full time data scientist in November 2019

Method overview

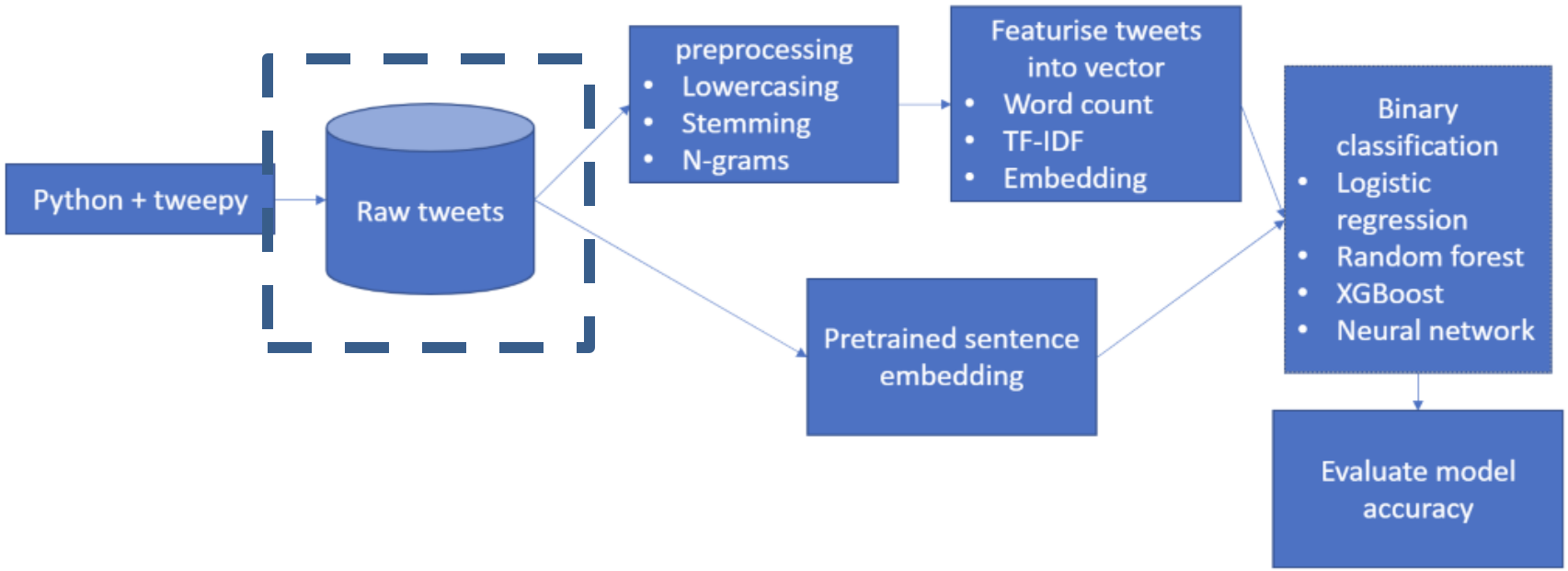


Data collection



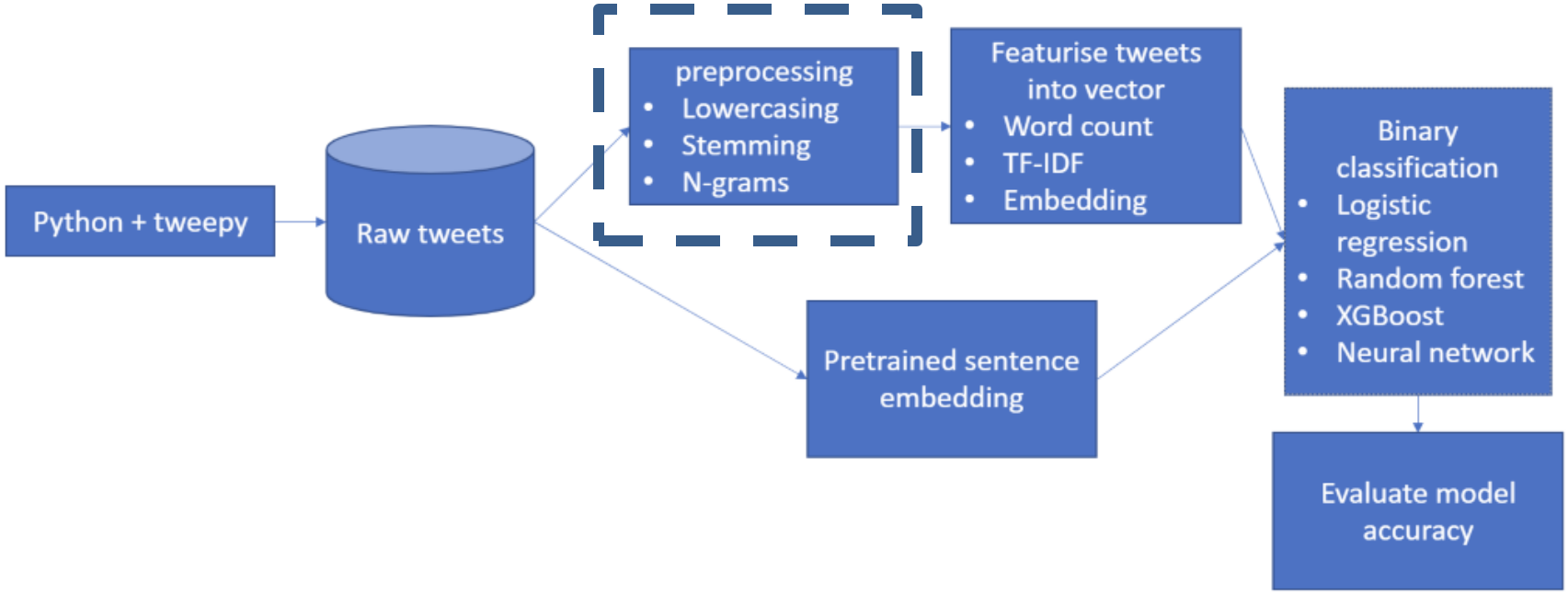
- Experimental data set
  - Only tweets containing happy or sad smileys --> supervised learning
  - Only Dutch tweets (contains tweets from other regions)

Data collection (2)



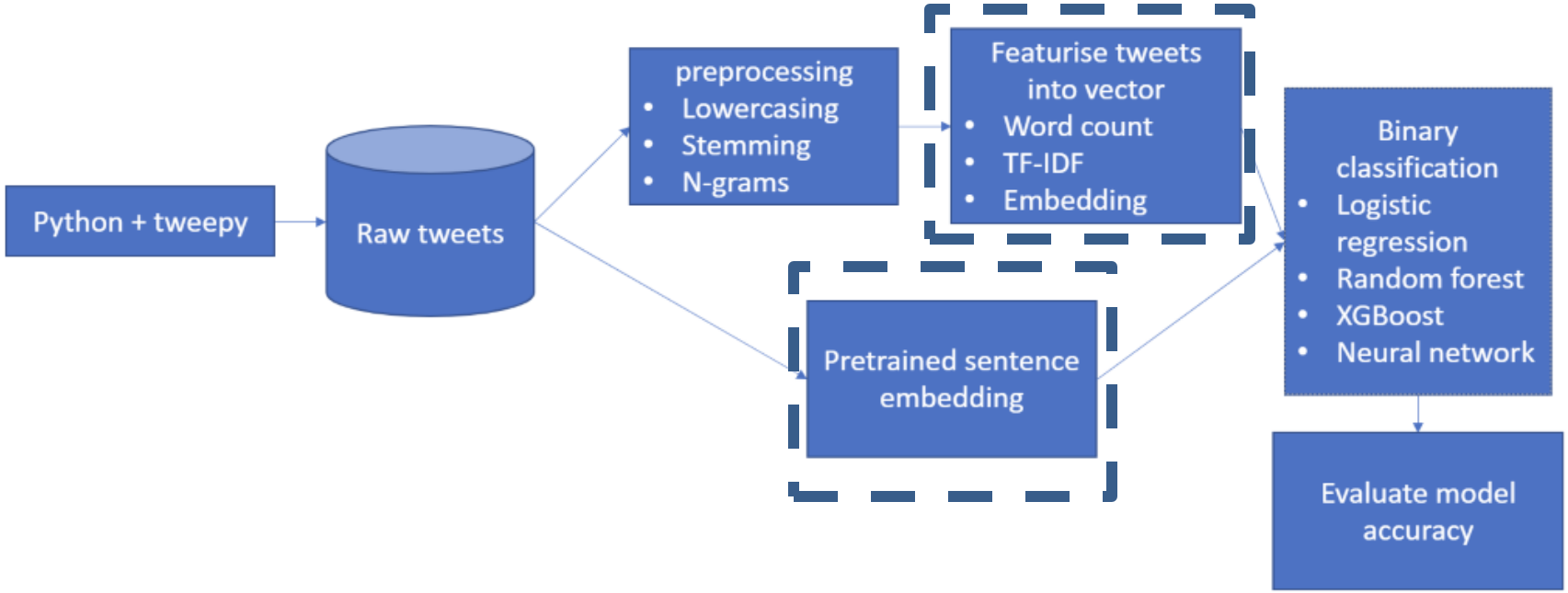
- Ran script once (data from 7 day period)
  - 19.000 Positive Tweets, 7.000 negative tweets
    - Remove smileys and add label

Data processing



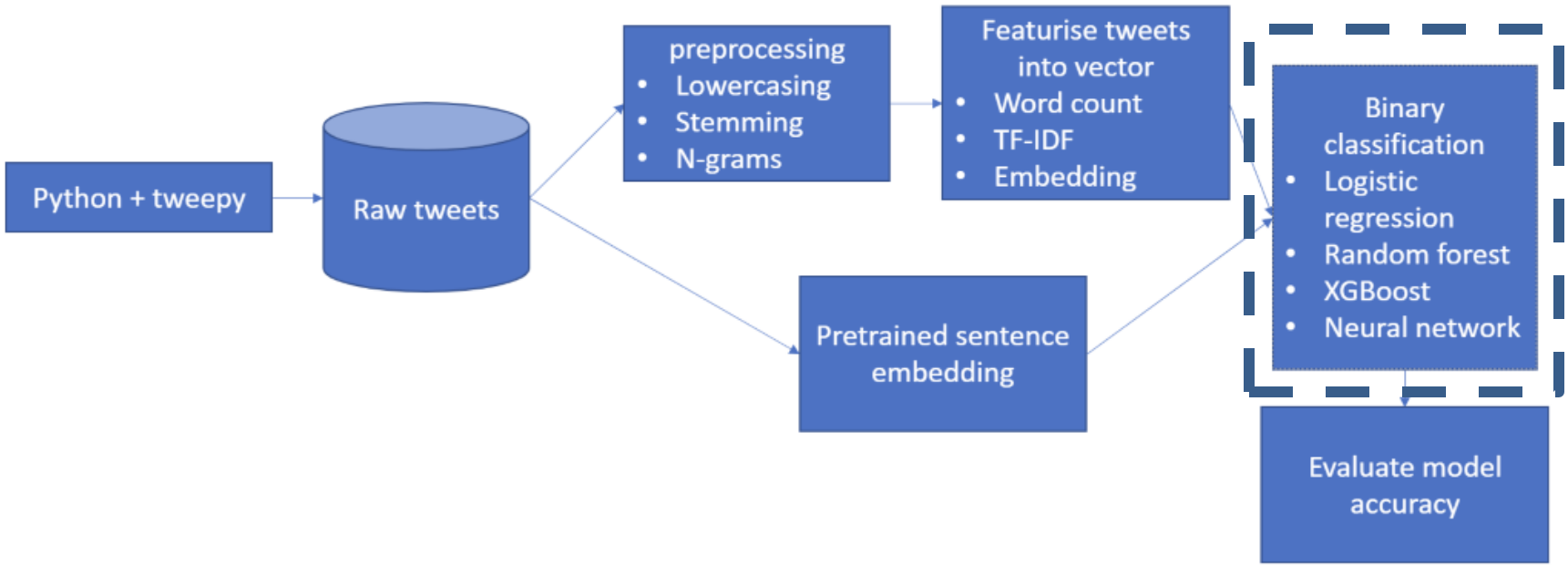
- Basic NLP preprocessing steps
  - Standardise text and reduce noise

Data processing (2)



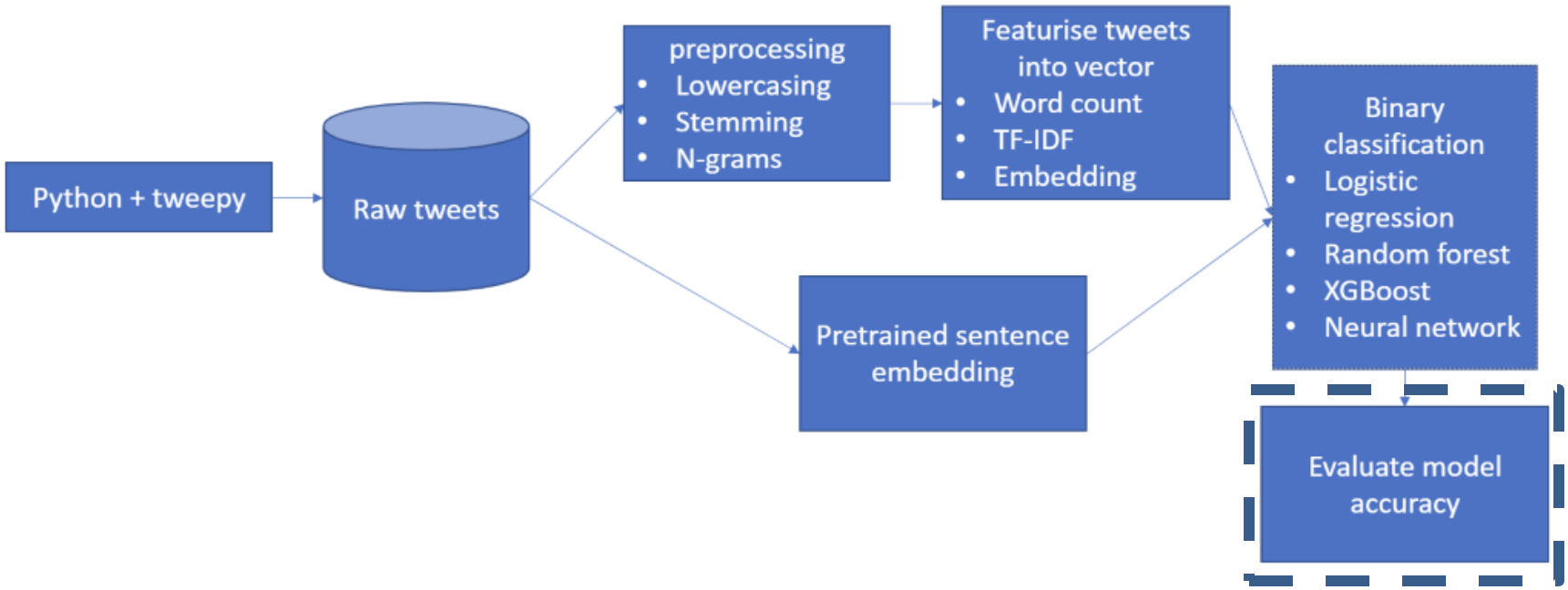
- From text to features
  - Evaluated different options
  - Word count performs best so far
  - Many more to try

Sentiment classification model



- From features to sentiment classification
  - Many candidate models evaluated (usual suspects)
  - Many more to try

Sentiment classification model (2)



- 5-fold cross validation on combinations of
  - Featurization algorithm
  - Classification algorithm
  - Hyperparameters of these algorithms

Sentiment classification model (3)

- Best model so far
  - Logistic regression in combination with count vector (many other combinations are close)
  - 81% accuracy

	precision	recall	f1-score	support
:(OR : (OR :'(	0.78	0.42	0.54	1810
:)	0.81	0.96	0.88	4822
micro avg	0.81	0.81	0.81	6632
macro avg	0.80	0.69	0.71	6632
weighted avg	0.80	0.81	0.79	6632

Future (and current) work

- Understand twitter data and biases
  - Lots of scepticism around this data source
  - Which part of Flemish population do we (not) get in the data?
  - Which part of all tweets do we get via the API?
- Continue work on model
  - Label data, so no more need for emoticons in text
  - Improve model performance
  - Q: how good is good enough?
- Set up continuous capturing of data

Conclusions

- Good experience so far
- Proven technical feasibility of a working twitter sentiment model
- Big questions remain before official statistics worthy
- A lot of ideas to improve on current state

Thank you for your attention

- Happy to answer any questions
- Grateful for any suggestions

- Contact: michael.reusens@vlaanderen.be

- Code (under construction)
  - https://github.com/mireusen/hlmos-statistiek-vlaanderen-twitter