# Finding the Smoke Signal: Smoking Status Extraction & Classification

ReMa Thesis "Linguistics and Communication Sciences"
Myrthe Reuver

## About this work and me:

- ResMA Linguistics & Communication Science,
  into clinical NLP. (One of two master's theses I wrote in 2020).

- Thesis written at Topicus, software company in Deventer
  making software for the public sector:
    - Data managing for general practitioners (GPs) in the Netherlands → data access;
    - GPs and Topicus' GDPR officer gave permission for this research: data was not allowed to
      be out of Topicus' servers.

- Since mid-September, PhD candidate at the Free University of Amsterdam (VU) on diversity in
  news recommender systems.

- Want to contact me for questions or ideas about my current or previous research?
  Email me:  myrthe[dot]reuver@vu[dot]nl

Radboud University

# Introduction: what is smoking status, and why extract it?

- Smoking status:
  - clinically relevant, often written in free text of GP (SOEP-text);
  - roughly 20% of Dutch adults smoke (CBS);
  - in NLP/clinical information science usually a task with 3 classes (smoker, ex-smoker, non-smoker) (Uzuner et. al., 2006)
  - or 5 classes (never, past, current, smoker temporally unknown, and unknown smoking status) (Wang et. al. 2019).
  - **high documentation load GPs**

- 'Care Standard' Tobacco Addiction 2019 (Trimbos Institute):

> **Onderdeel 1 – Adviseren**
> Stel vast of de patiënt rookt door middel van de vraag: "Rookt u wel eens? Doet u dit dagelijks of af en toe?"
> Adviseer elke patiënt die (weleens) rookt om te stoppen met roken, waarbij u het advies toespitst op de situatie van de patiënt en informatie geeft over effectieve behandelmogelijkheden

topicus

## The task

Predict, based on the free text in an EMR, the smoking status of a primary care patient.

"Meneer zegt niet te roken"  →  **smoking status: no smoker**
(mr. says he does not smoke)

- Idea: used **supervised learning** (teach a model with examples)

# Problem in Earlier Smoking Status Classification Work

- Small training sets e.g. Uzuner (2006) → 502 EMRs, Weng et. al. (2019) --> 475 EMRs → especially not enough training examples for neural models

- Sparsely labelled → roughly 2% of the Electronic Medical Records (EMRs)'s consultations has a recorded smoking status in our dataset

- Mainly tested on 'clean' benchmarking datasets in the literature (ib2b 2006 shared task, Mayo Clinic dataset in Wang et. al. (2019):
    - **pro:** open data
    - **con:** not realistic in real clinical settings, sparsely labelled data

**Our goal: overcome this sparsely labelled data problem and improve over simple, rule-based models, on 'real' clinical data.**

**"How can we best automatically detect and classify the smoking status in primary care patients' EMR on the basis of the free text in GP doctor's notes, and overcome the sparsely labelled data problem?"**

topicus

# Important ethical and methodological concerns

Topicus was interested in finding all unknown smoking statuses in EMRs.

*However:*

We are classifying DOCUMENTS, not people.
These documents ≠ consistent or reliable representation of real-world people.

Also, some inherent biases leads to imperfect detection;

- a positive smoking status will more often be recorded, leading to less detection of non or ex-smokers;
- doctors will record the smoking status of certain patient groups more (e.g. chronic illness), leading to any model's bias towards detecting smoking status in this group;
- **absence of any mention of smoking in EMR does not automatically mean non-smoking for a patient!**

Radboud University

topicus

# LITERATURE ON PROBLEM AND METHODS

# Smoking status extraction & classification → working with sparsely labelled data

- Kreimeyer (2019), systematic literature review: **46% of clinical NLP projects** aiming to identify and extract elements from unstructured text in EMRs still **use rule-based systems**,

- **A. Rule-based → regular expressions** → used before in Weng 2019, Palmer 2019, Uzuner 2006, reporting over **90% accuracy (!)**
  - **pro:** this problem has relatively fixed vocabulary ("tabak", "nicotin*")
  - **con:** not very flexible, **cannot detect patterns not noted by rule designers**

- **B. Increasing training data → weak supervision → SNORKEL**
  - **pro:** works with rules, which works well with this problem, while also able to use training data in a machine learning model
  - Wang et. al. (2019) claims to use it, but their paper only gives evidence of simple rule-based labelling (?)

- **C. Transfer learning → BERT → fine-tuning**
  - **pro:** language model already retains semantic information useful for classification;

# SNORKEL (Ratner et. al. 2017)
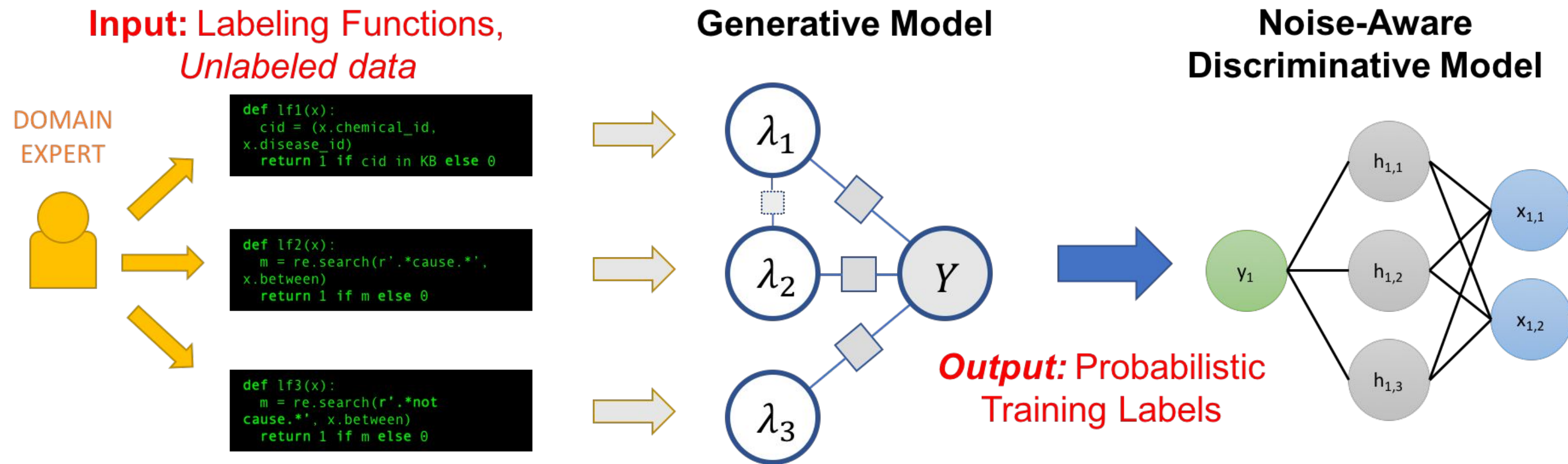


- works with Labelling Functions (LFs), heuristics or rule-based labellers
- These can be optimized on a small labelled **development set**

- LFs are **weighted** in a LabelModel
- exploiting (dis)agreements between LFs → each LF as an independent labeller ("Wisdom of the crowds")

Radboud University

topicus

# BERT & BERTje

- BERT (Devlin et. al. 2019): large-scale, pre-trained transformer trained on a **masking** task: predicting context from words.
- In this manner, **semantic information can be retained,** useful for newer tasks

- We use BERTje (de Vries 2019), 12 layer Transformer model trained on Dutch Wikipedia, SoNaR, and other data in a masking and sentence prediction task.
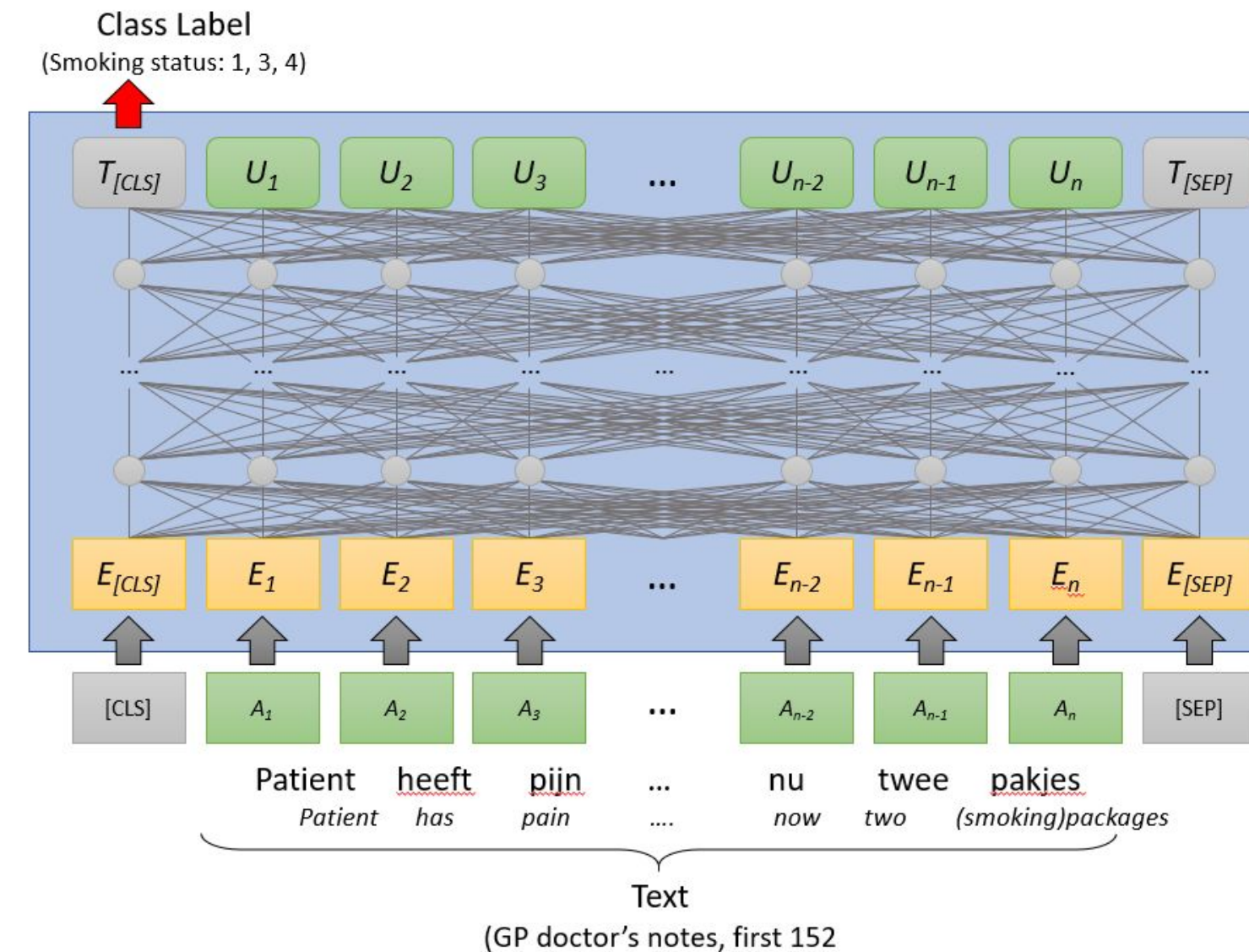
**What is a vector representation? Can represent words, but also documents → similar contexts = similar vectors**
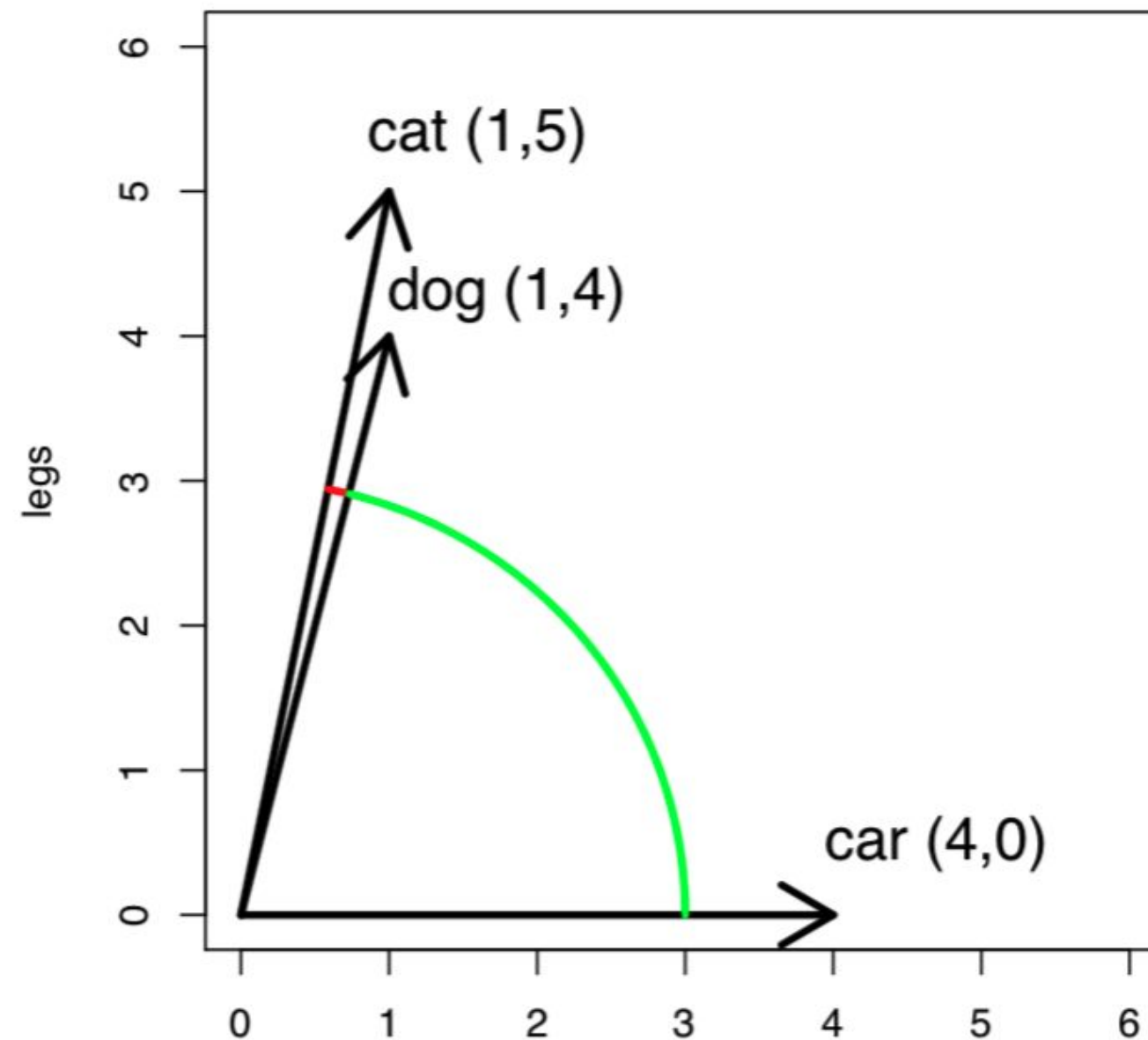


from Baroni & Boleda

# DATA

# Data & Preprocessing OF EMRs:

- 6 GP offices in the Netherlands.
- Each GP office has 4 datafiles: **PATIENTS, EPISODES, MEASUREMENTS** (many different smoking variables), **CONSULTATIONS (**SOEP-text).
- 943.757 **consultations** in 24 data files

Our preprocessing:
**1. combining** and **filtering** these 24 datafiles into one datafile

**2. Normalization of EMR:** only the **last** consultation for each patient.
Smoking status: P1739 → 3 classes: EX-SMOKER, SMOKER, NEVER.

**3. Filtering out duplicates and minors**

Final dataset:  **17.873 EMR representations**

**4. Data split:** train (80%), dev (10%), and test (10%) split

Radboud University

topicus

# EMR representation

| patient ID_GP ID | Sex | Age at consult | Age in 2020 | SOEP text | date | smoking (1739) | Ketenzorg |
|---|---|---|---|---|---|---|---|
| 9999_777 | F | 40 | 43 | Mevrouw heeft buikpijn<br><br>Translation: Mrs. has stomach pain | 23-04-2017 | 4 | 0 |
| 8888_666 | M | 63 | 62 | Is gestopt met pasta eten, is afgevallen.<br><br>Has stopped eating pasta, has lost weight | 05-07-2019 | 1 | 1 |

# Dataset: size and labelled sub-set

| | Training set | Development | Test |
|---|---|---|---|
| EMR representations | 14.298 | 1.788 | 1.787 |

Table 6: The **labelled** datapoints in values used for the smoking status variable '1739' ("smoking"), as defined by the NHG (National GP Association)

| | Training | Dev | Test |
|---|---|---|---|
| "smoker" | 794 | 115 | 103 |
| "never smoked" | 2081 | 268 | 274 |
| "ex-smoker" | 2103 | 268 | 251 |
| total labelled EMR representations | 4.978 | 651 | 628 |

# METHODS
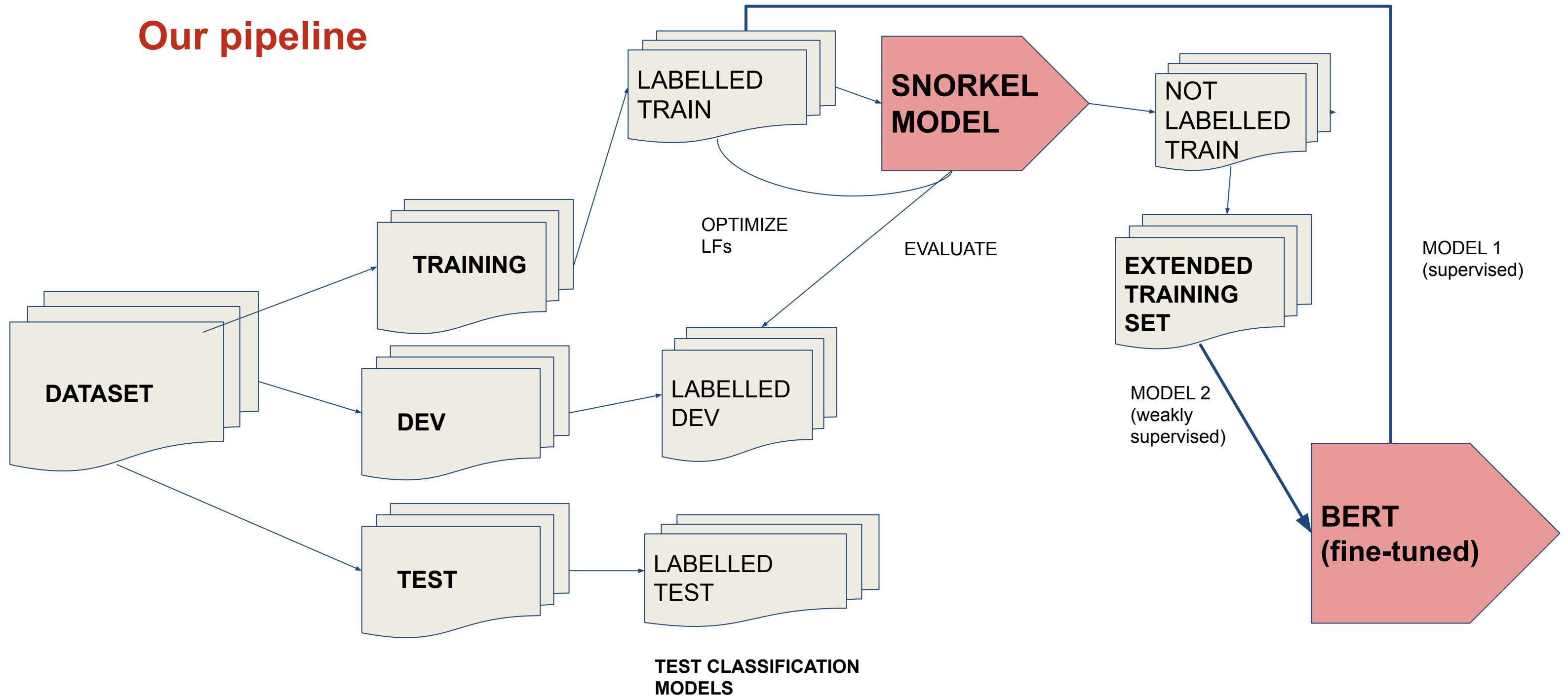
# Our comparison in smoking status classification

**Compare:**

- rule-based baselines (based on earlier work + Care Standard);
- BERTje;
- SNORKEL + BERTje (larger training set).

**Evaluation:**

- precision, recall, F1 → do we correctly predict smoking status?
- confusion matrices → When we incorrectly predict, what does the model predict?

# Transfer Learning with BERTje: fine-tuning

Training process:

- first: tokenize dataset with BERTje tokenizer;

- add one linear layer to BERTje, predicting 3 classes (smoker, non-smoker, ex-smoker)

- training: 3 epochs, learning rate: 0.00005
  → more epochs = overfitting (training loss lower than development loss)

# Weak Supervision with SNORKEL - LFs

- Started with 32 heuristics, mostly based on keywords based on earlier literature and the Zorgstandaard:

```
"""rookt --> third person present."""
keyword_rookt = make_keyword_lf(keywords=["rookt"], label=SMOKER)

"""roker --> noun smoker."""
keyword_roker = make_keyword_lf(keywords=["roker"], label=SMOKER)

"""was smoker --> past."""
keyword_roker_was = make_keyword_lf(keywords=["was roker"], label=EX)
```

- LabelModel trained with 500 epochs, learning rate 0.01

topicus
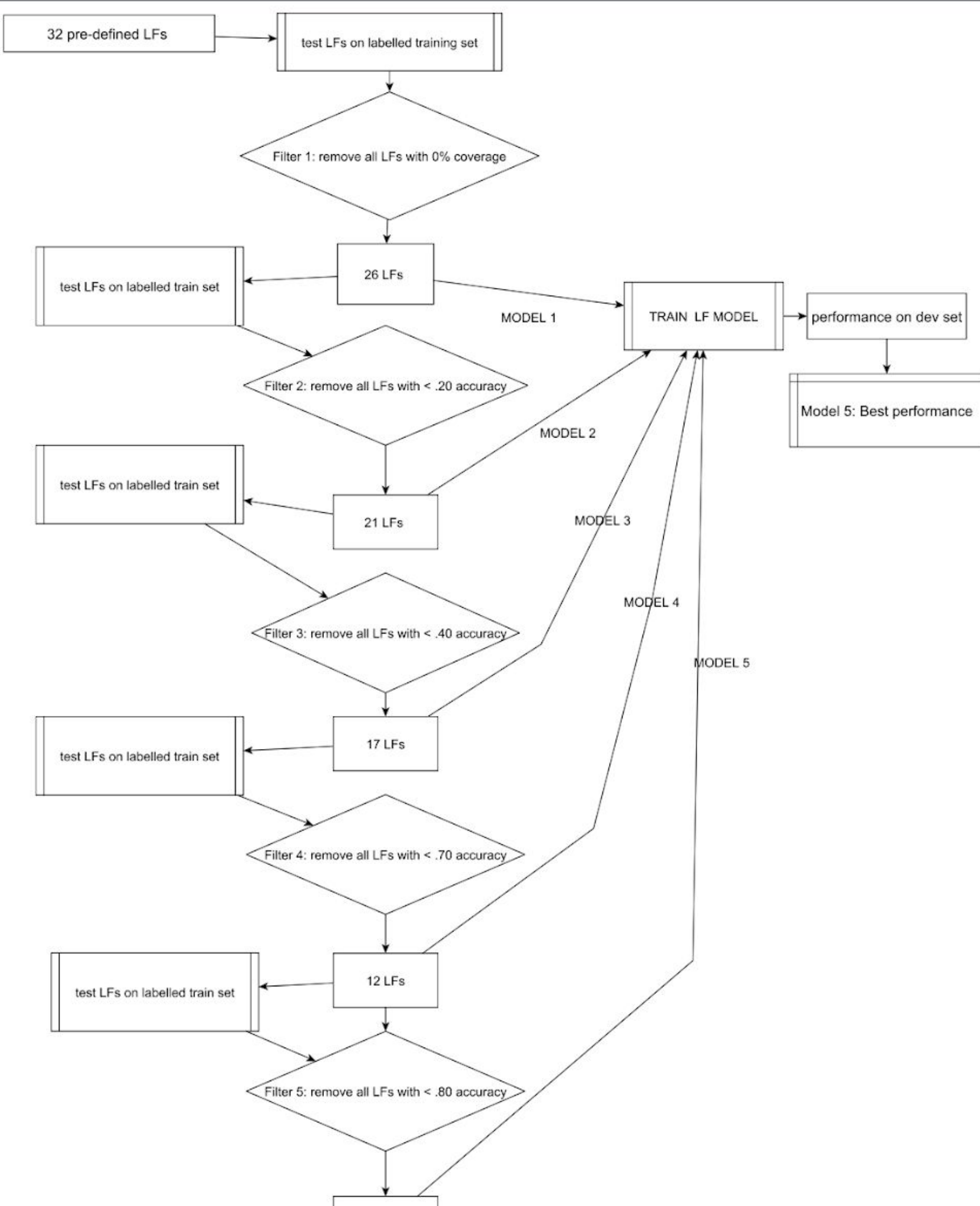
# SNORKEL: training a LabelModel

Interesting results LFs:
- of all 'quit smoking' medicines mentioned in the health directive, only "champix" had any coverage;

- "roken" gave opposite result: the word was more often mentioned with people who never smoked (.45 accuracy) than with smokers, which was expected (.19 accuracy).
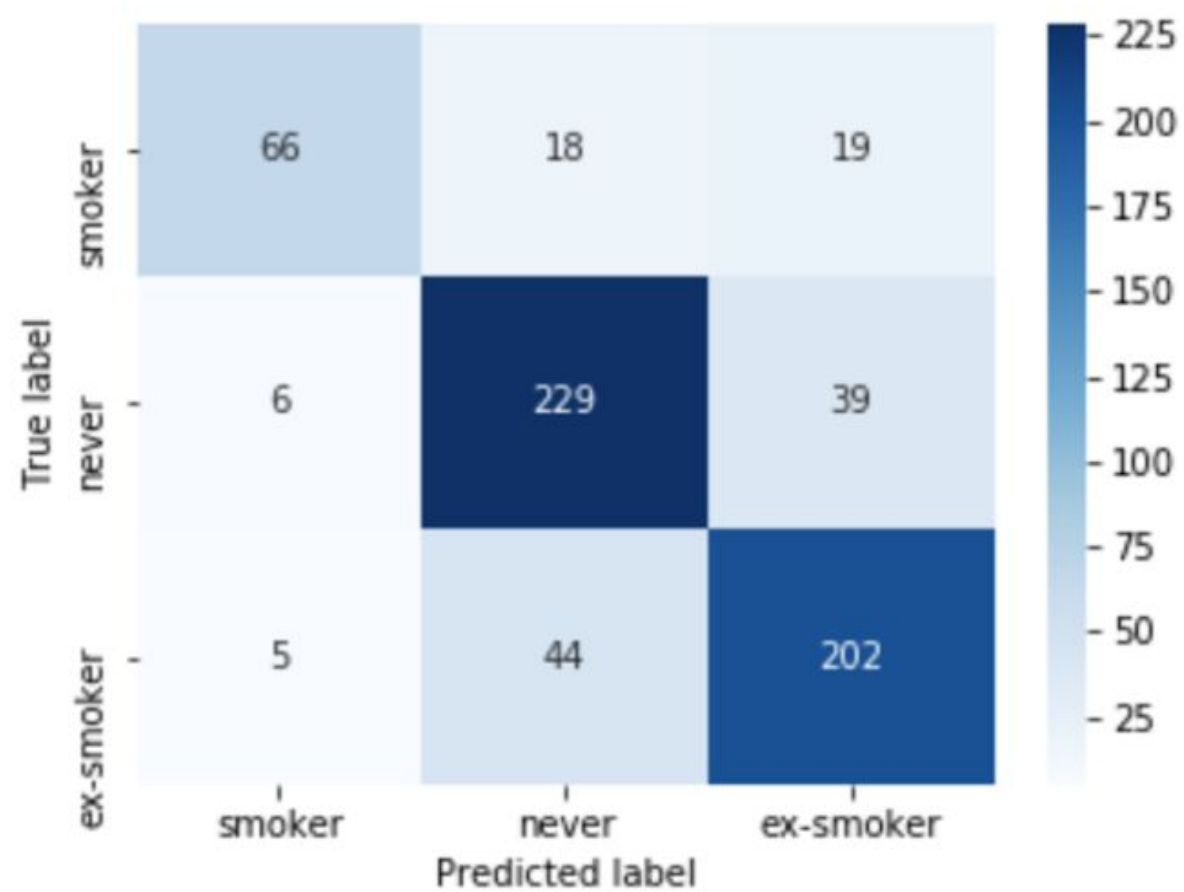
snorkel

Radboud University

topicus

**RESULTS**

# Results on the test set: overall and in-class

|  | Rule-Based | BERTje | SNORKEL + BERTje |
|---|---|---|---|
| precision (micro) | 0.49 | **0.79** | **0.79** |
| recall (micro) | 0.43 | **0.79** | **0.79** |
| F1 (micro) | 0.55 | **0.79** | **0.79** |

|  | BERTje *4.978 training examples* | | | SNORKEL+BERTje *5.490 training examples* | | |
|---|---|---|---|---|---|---|
|  | precision | recall | F1 | precision | recall | F1 |
| SMOKING | 0.82 | 0.64 | 0.72 | **0.86** | 0.64 | **0.73** |
| NON-SMOKING | 0.74 | 0.76 | 0.75 | **0.79** | **0.84** | **0.81** |
| EX-SMOKING | **0.82** | **0.83** | **0.82** | 0.78 | 0.80 | 0.79 |

# Confusion Matrices (on the Test set)

BERTje

BERTje + SNORKEL:
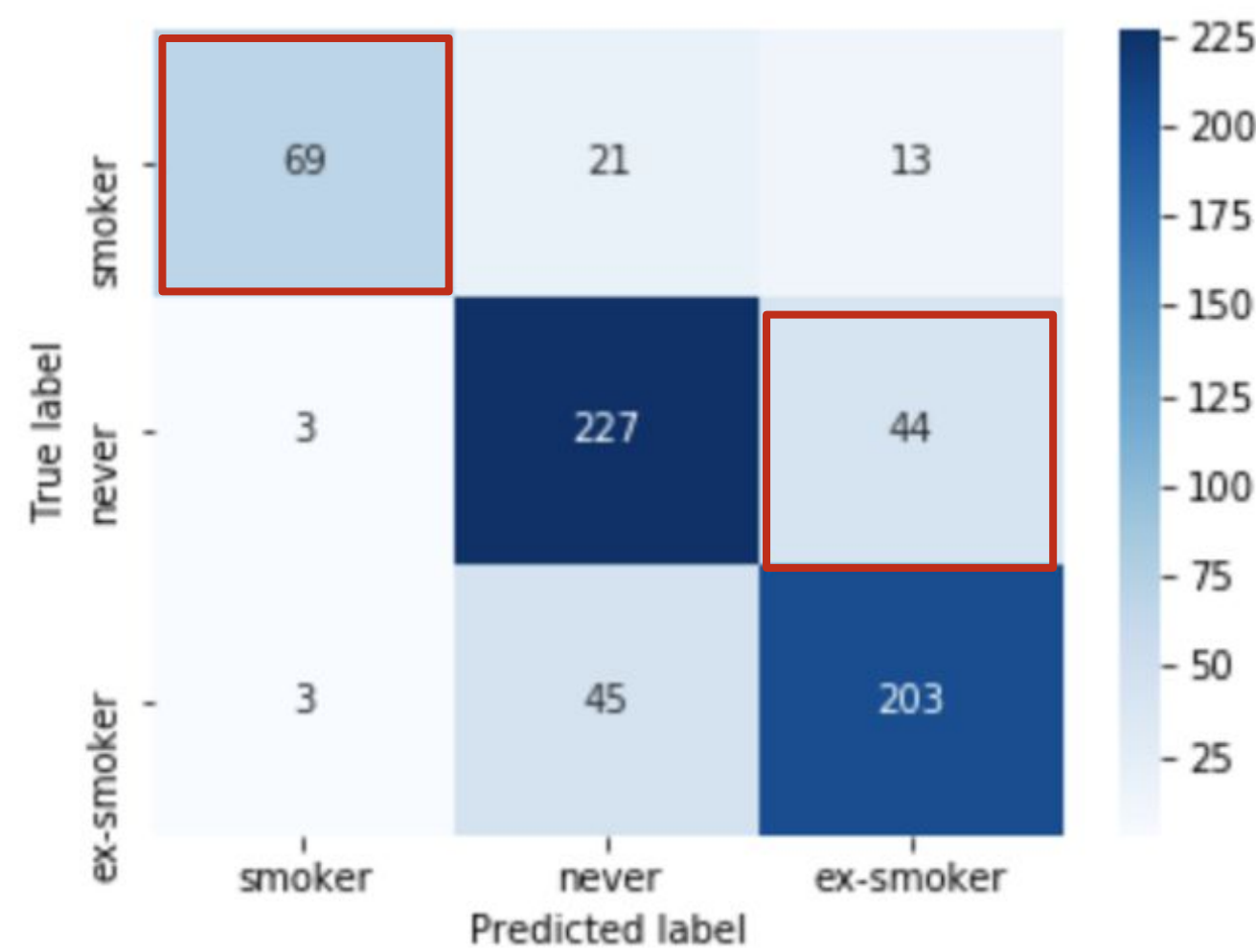
# CONCLUSION

## Things we learned

○ real-world data is more complicated than shared task data

**"How can we best automatically detect and classify the smoking status of primary care patients' EMR on the basis of the free text in GP doctor's notes, and overcome the sparsely labelled data problem?"**

● Weakly supervised method works for some classes (SMOKING, NON-SMOKING), where there is in-class improvement, but no overall improvement over supervised learning;

● Rule-based method → does not seem to generalize well;

● A model trained on general language understanding (BERTje) is surprisingly not very bad at smoking status classification.

topicus

# DISCUSSION

# Research idea / theorizing / neat set-up and preprocessing / results
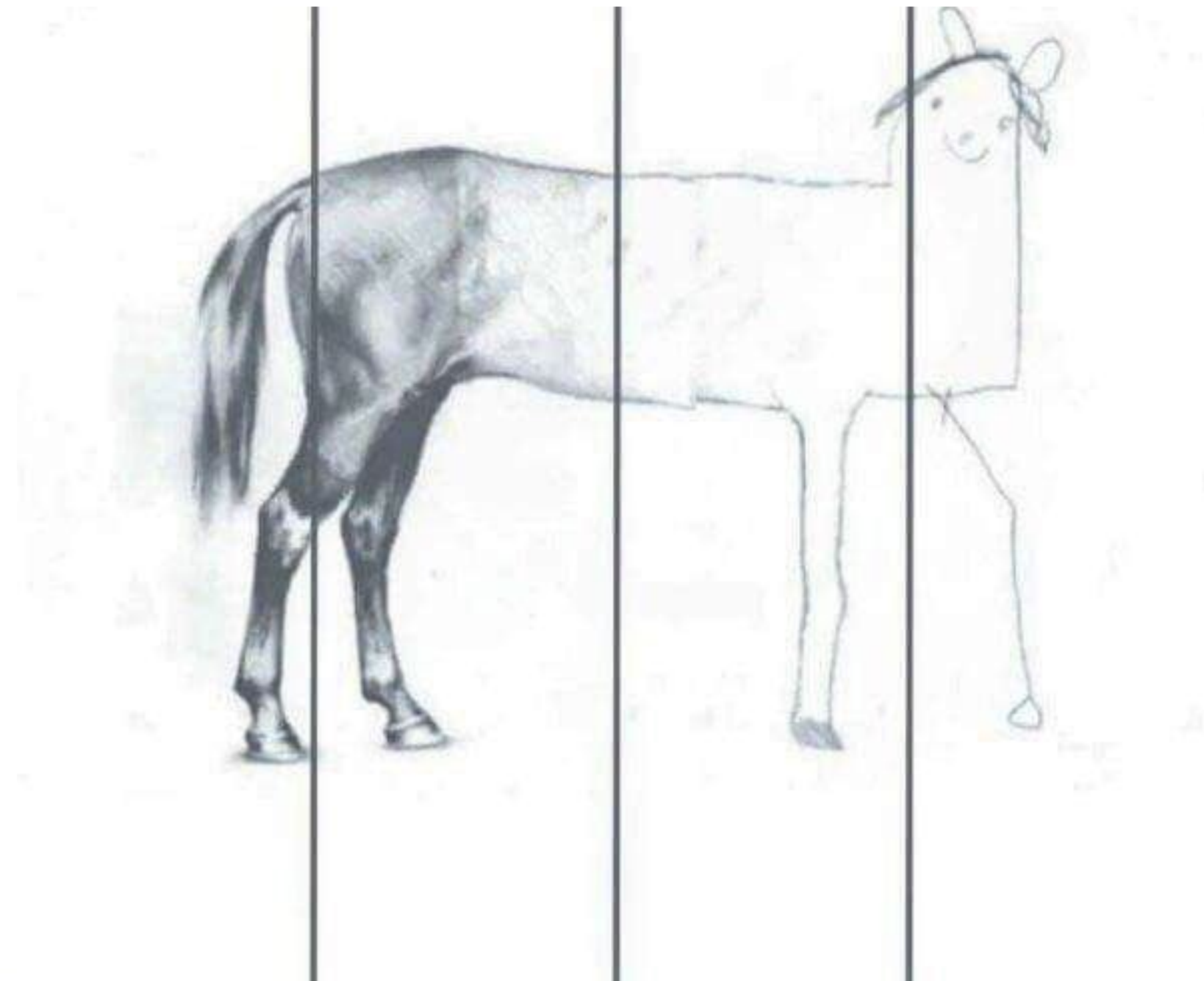
# References

Uzuner, Ö., Goldstein, I., Luo, Y., & Kohane, I. (2008). Identifying patient smoking status from medical discharge records. Journal of the American Medical Informatics Association, 15(1), 14-24.

Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., ... & Liu, H. (2019). A clinical text classification paradigm using weak supervision and deep representation. BMC medical informatics and decision making, 19(1), 1.

CBS. 2018.  Helft van laagopgeleide 25- tot 45-jarige mannen rookt.
https://www.cbs.nl/nl-nl/nieuws/2018/22/helft-van-laagopgeleide-25-tot-45-jarige-mannen-rookt

Ford, E., Carroll, J. A., Smith, H. E., Scott, D., & Cassell, J. A. (2016). Extracting information from the text of electronic medical records to improve case detection: a systematic review. Journal of the American Medical Informatics Association, 23(5), 1007-1015.

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., ... & Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. Journal of biomedical informatics, 73, 14-29.

Palmer, E. L., Hassanpour, S., Higgins, J., Doherty, J. A., & Onega, T. (2019). Building a tobacco user registry by extracting multiple smoking behaviors from clinical notes. BMC medical informatics and decision making, 19(1), 141.

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). Bertje: A dutch bert model. arXiv preprint arXiv:1912.09582.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017, November). Snorkel: Rapid training data creation with weak supervision. In Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases (Vol. 11, No. 3, p. 269). NIH Public Access.

Radboud University

topicus