Active Learning for Filtering

UvA and UNED at RepLab 2013

Maria-Hendrike Peetz, Damiano Spina, Maarten de Rijke, Julio Gonzalo

Combine expert knowledge with computational effort by asking the right questions.

Task: Identify tweets that are relevant to a company.

Annotating few examples is ok.

1. Analysts need to keep up to date anyway.
2. They have the ultimate responsibility for the result.

Data changes over time.

@nokia: My phone stopped working.

2012

@microsoft

2013

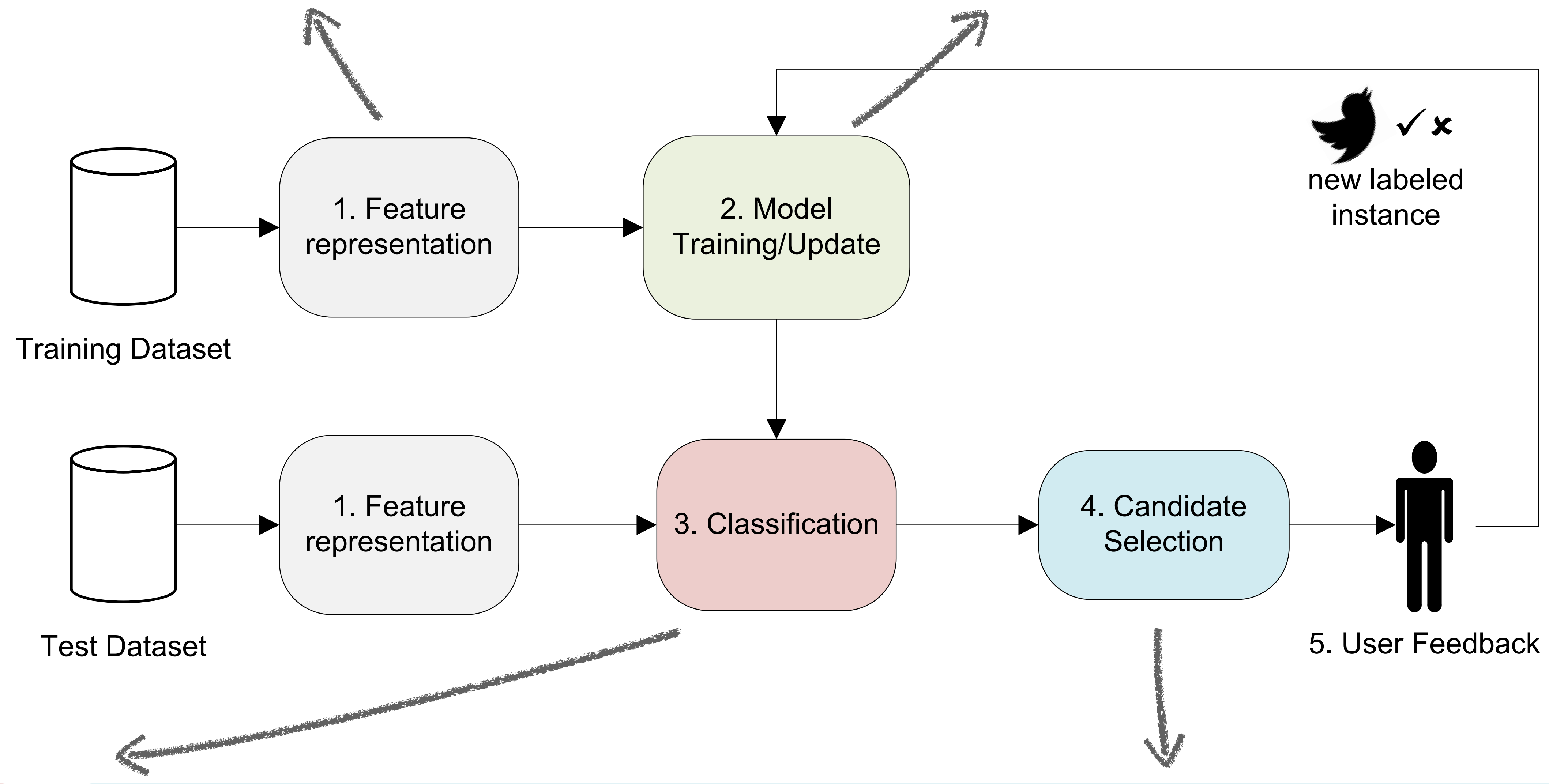
Sample tweets, annotate them, and update the model.

Feature representation

- Tweet metadata (always)
- Entity linking of the tweets (BoE)

Model Training/Update

- Retrain NB with every new instance.
- Higher weight to newly annotated instances.



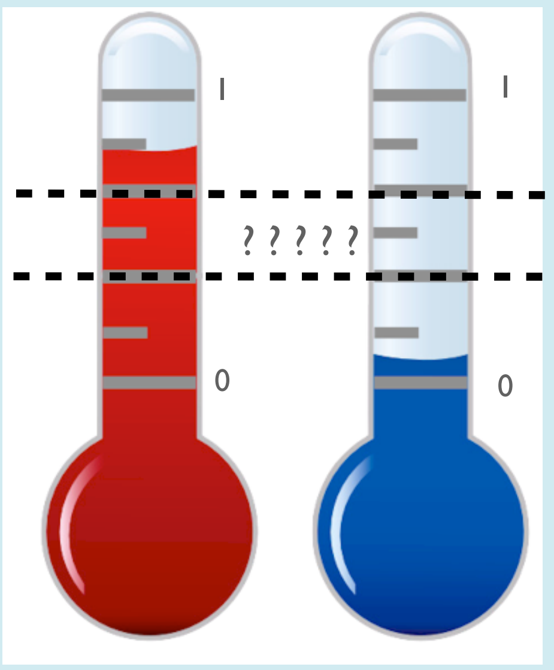
Classification

- Naive Bayes

Candidate Selection

Margin/uncertainty sampling:

Selecting examples close to the margin means sampling examples where the classifier is less confident.



Select 1% of tweets for annotation:

On average that are 15 tweets per entity, in the language dependent case: 10 English, 5 Spanish

With a decent baseline, annotating 15 tweets per entity is enough.

Results

run id	accuracy	R	S	F(R,S)
Jaccard	0.8714	0.4902	0.3200	0.3255
BoE+lang. independent	0.2785	0.1635	0.1258	0.0730
BoE+lang. dependent	0.2847	0.2050	0.1441	0.0928
BoE+lang. independent + AL	0.5657	0.2040	0.2369	0.1449
BoE+lang. dependent + AL	0.6360	0.2386	0.2782	0.1857
BoE+lang. independent + AL + majority	0.7745	0.6486	0.1833	0.1737
BoE+lang. dependent + AL + majority	0.8155	0.6780	0.2187	0.2083

Lessons Learnt

- Filtering models worked well on the trial data
- Active learning with 1% improves weak classifiers
- There was not enough training data for language dependent training.

ISLA, Informatics Institute
University of Amsterdam
http://ilps.science.uva.nl

LiMoSINe

M.H.Peetz@uva.nl

COMMIT/

damiano@lsi.uned.es

NLP & IR Group
UNED
http://nlp.uned.es

UNED