Measuring online sentiment with BERT

Pieter Delobelle

Nov 24, 2023

DTAI
Declarative Languages & Artificial Intelligence

KU LEUVEN

Outline

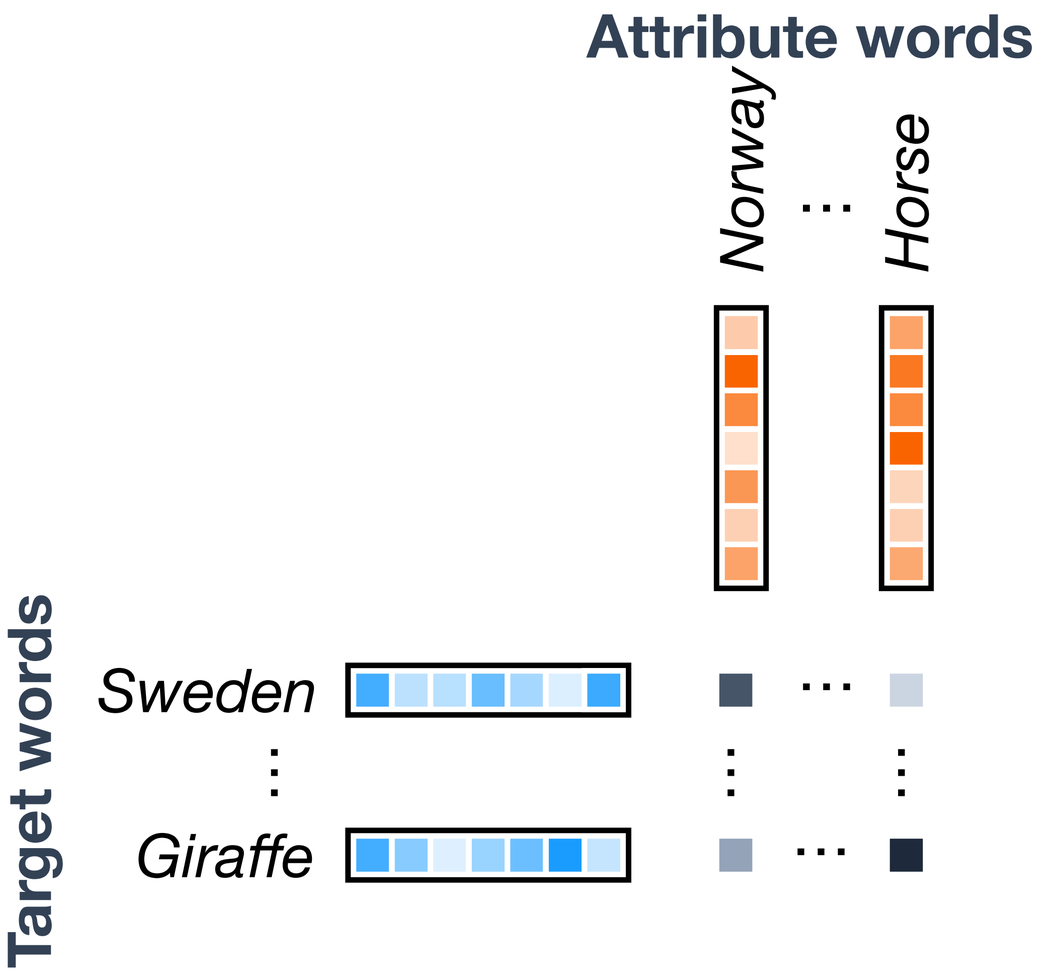
- **Language models:** BERT, GPT, ...
  - What?
  - How?
  - Where?
- **Measuring sentiment**: illustration with COVID measures
- **Topic modelling**
- **Practical session**

Language models

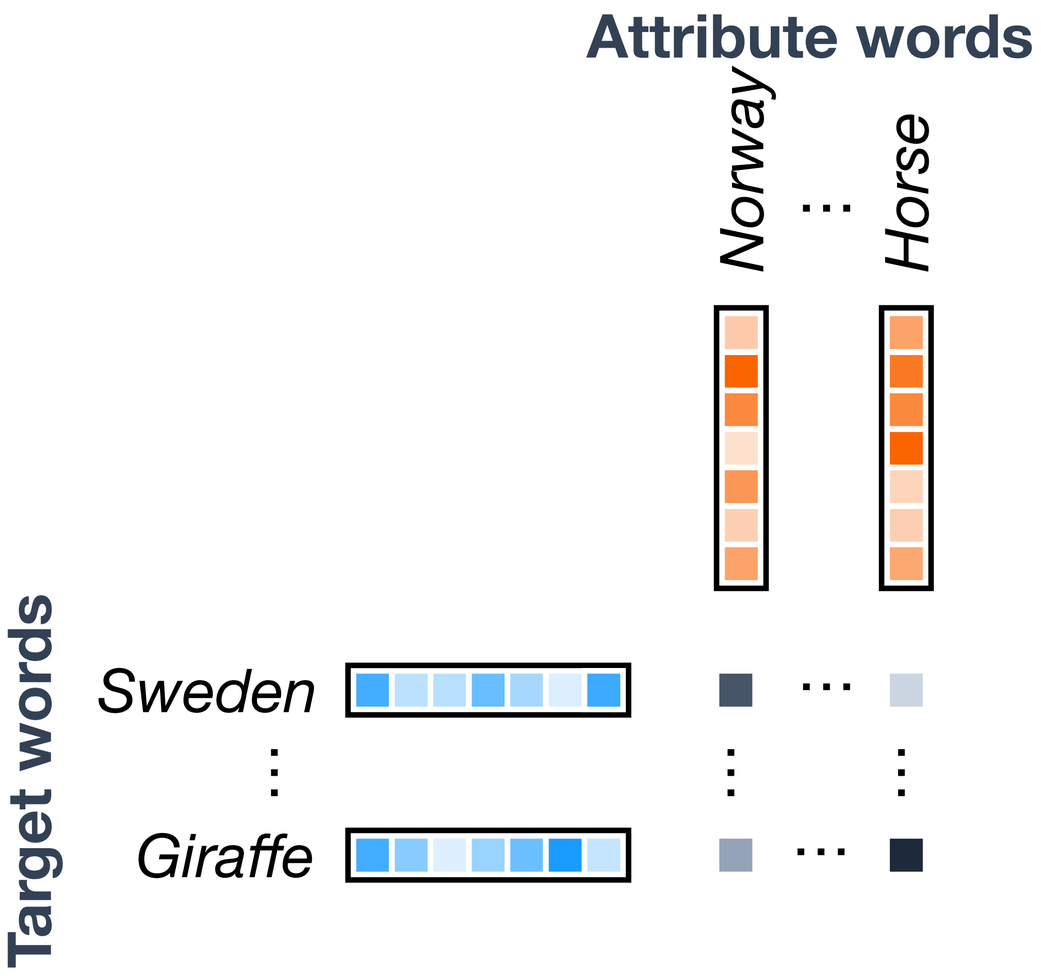
Word embeddings

stick

Word embeddings

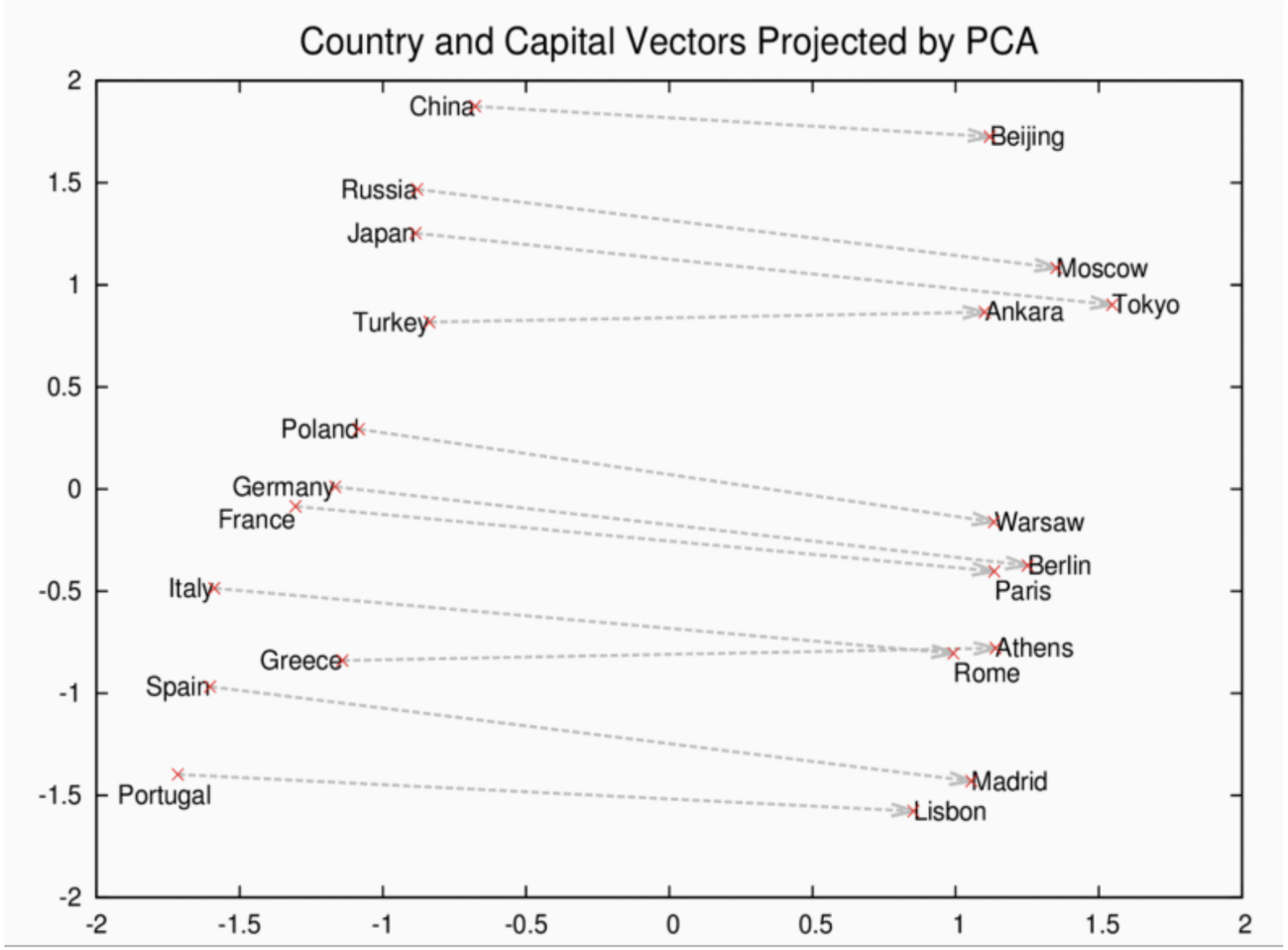


Word embeddings



Word	Cosine distance
norway	0.760124
denmark	0.715460
finland	0.620022
switzerland	0.588132
belgium	0.585835
netherlands	0.574631
iceland	0.562368
estonia	0.547621
slovenia	0.531408

Embeddings have meaningful principal components



https://wiki.pathmind.com/word2vec

Word embeddings

stick

Word embeddings don’t understand polysemy

Bank

Bank

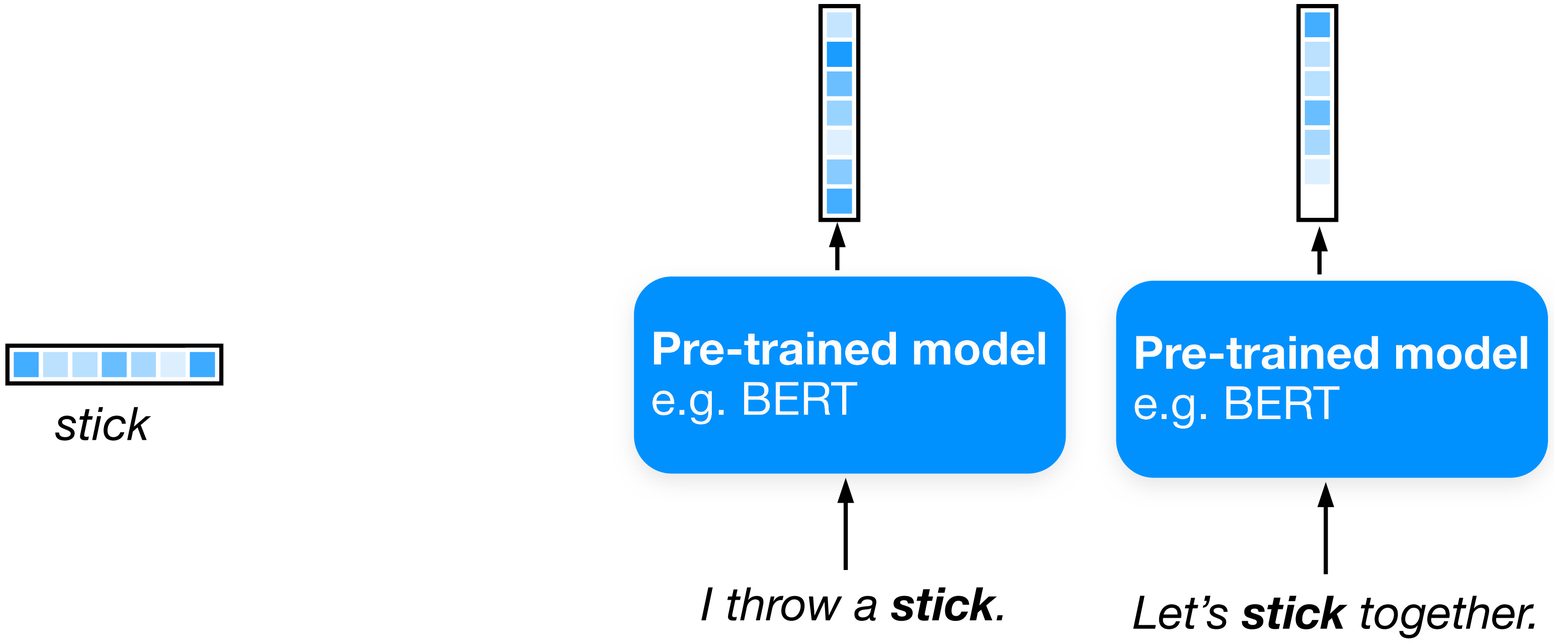
Word embeddings don’t understand polysemy

Bank

Bank

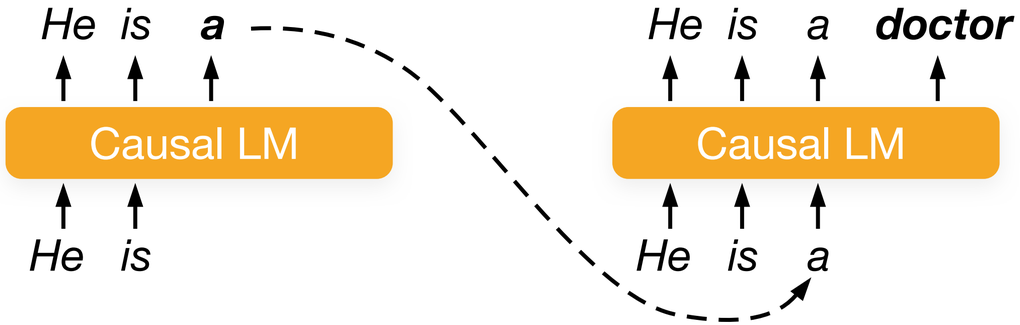
➜ How to incorporate context?

Language models address polysemy

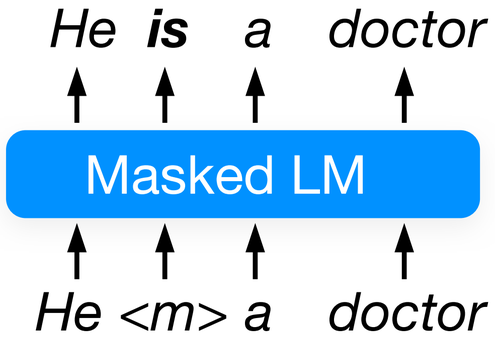


Language modeling

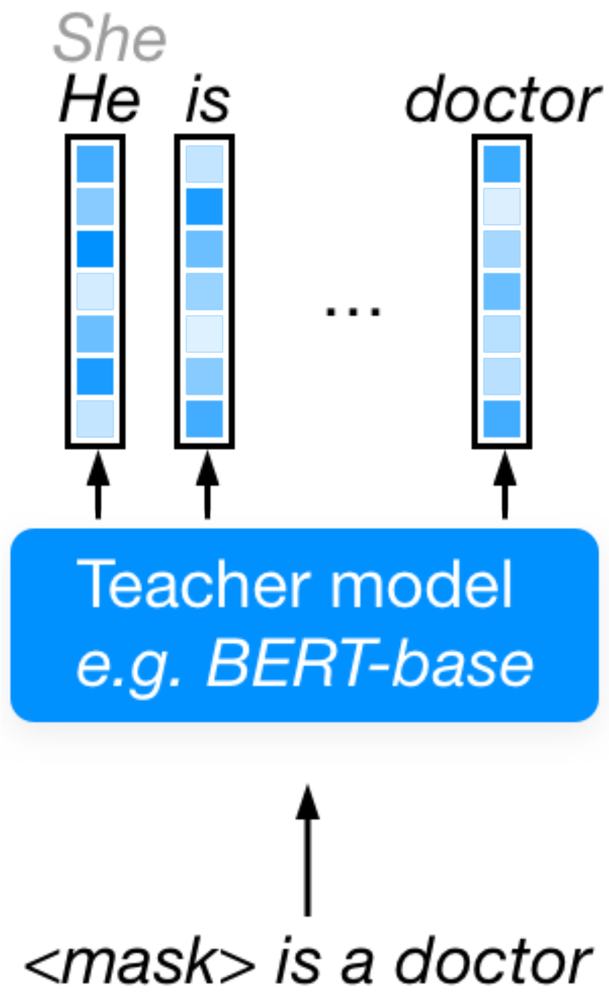
1. Causal language modeling (CLM)



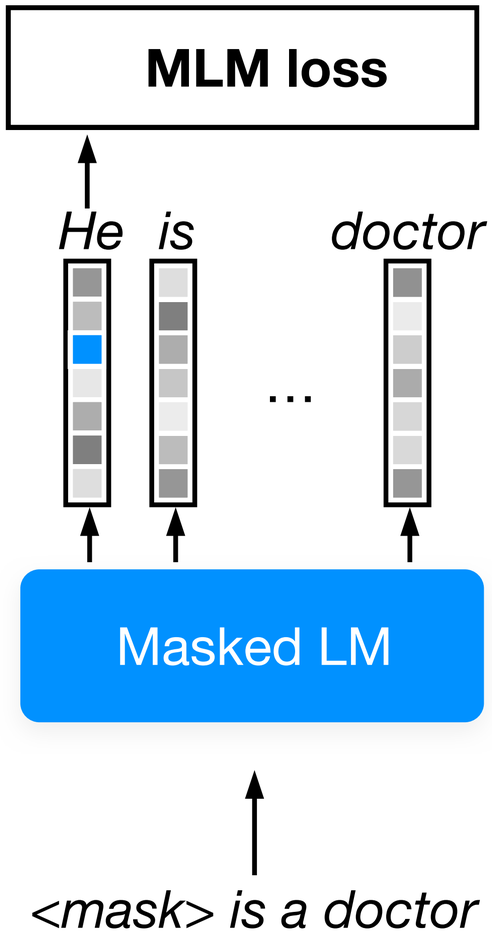
2. Masked language modeling (MLM)



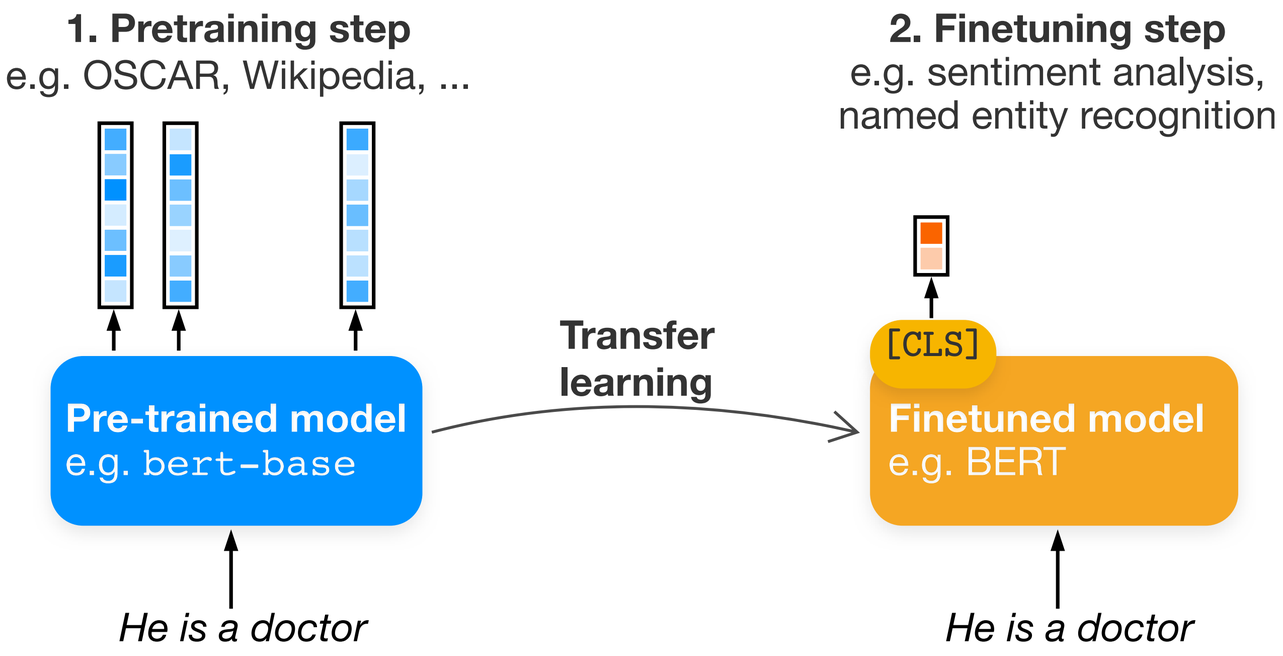
MLMs learn a probability for each word



MLMs learn a probability for each word



MLMs are trained twice



Where? On the HuggingFace hub

Models 408,066 | Filter by name | new Full-text search | ↑↓ Sort: Trending

- stabilityai/stable-video-diffusion-img2vid-xt
  Updated 2 days ago • 628
- openai/whisper-large-v3
  Automatic Speech Recognition • Updated 2 days ago • 99k • 882
- microsoft/Orca-2-13b
  Text Generation • Updated 1 day ago • 2.77k • 289
- stabilityai/stable-video-diffusion-img2vid
  Updated about 16 hours ago • 181
- coqui/XTTS-v2
  Text-to-Speech • Updated 6 days ago • 66.4k • 292
- Intel/neural-chat-7b-v3-1
  Text Generation • Updated about 7 hours ago • 4.84k • 152
- openchat/openchat_3.5
  Text Generation • Updated 1 day ago • 24.9k • 674
- latent-consistency/lcm-lora-sdxl
  Text-to-Image • Updated 8 days ago • 51.5k • 360
- google/switch-c-2048
  Text2Text Generation • Updated 4 days ago • 595 • 159
- 01-ai/Yi-34B
  Text Generation • Updated about 5 hours ago • 58.7k • 993
- mistralai/Mistral-7B-v0.1
  Text Generation • Updated Oct 12 • 436k • 1.93k
- microsoft/Orca-2-7b
  Text Generation • Updated 1 day ago • 2.22k • 97
- HuggingFaceH4/zephyr-7b-beta
  Text Generation • Updated 1 day ago • 169k • 883
- meta-llama/Llama-2-7b-chat-hf
  Text Generation • Updated 11 days ago • 950k • 1.89k
- stabilityai/stable-diffusion-xl-base-1.0
  Text-to-Image • Updated 25 days ago • 9.97M • 3.59k
- teknium/OpenHermes-2.5-Mistral-7B
  Text Generation • Updated 20 days ago • 16.7k • 200

https://huggingface.co/models

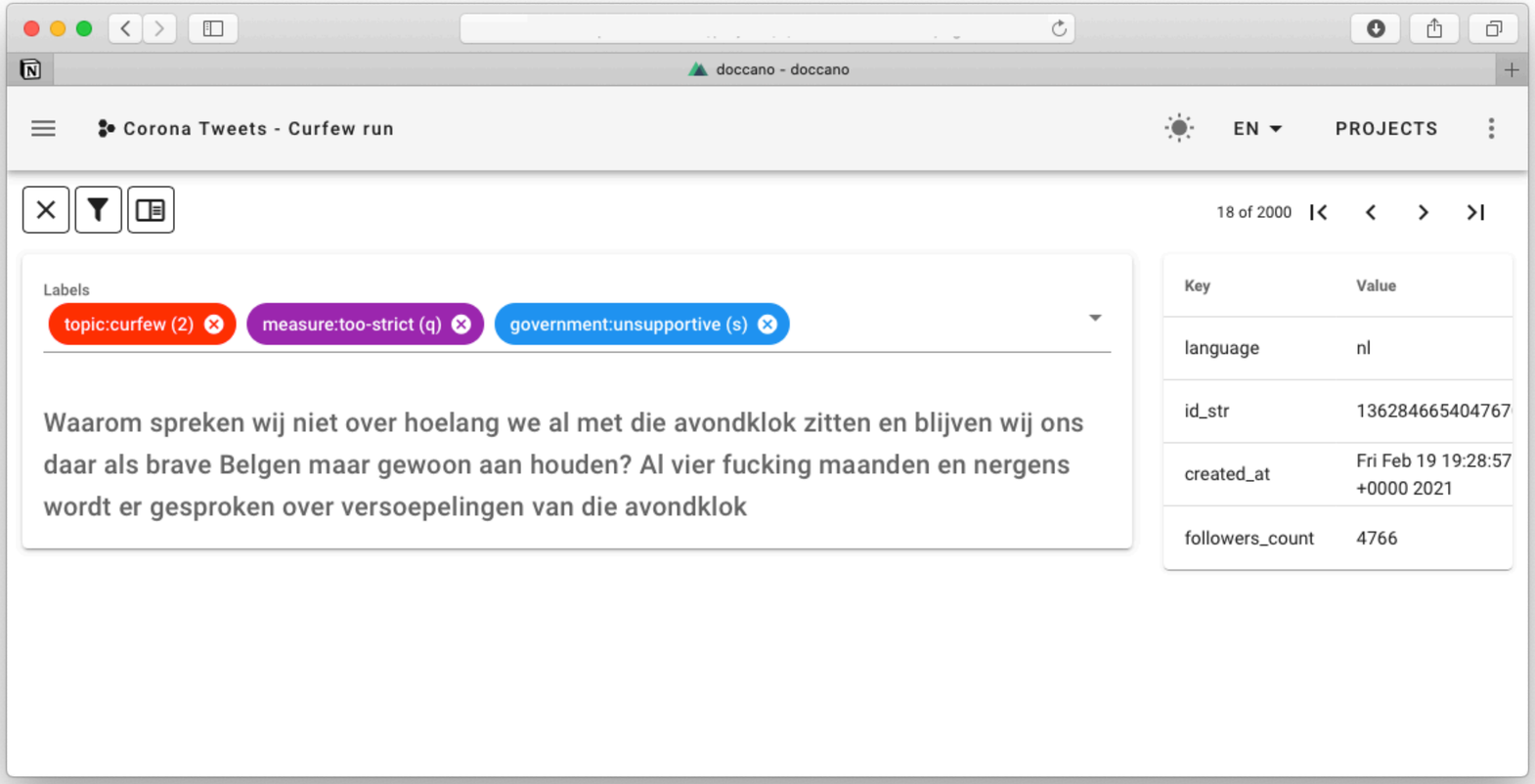
Measuring sentiment on Twitter

Context

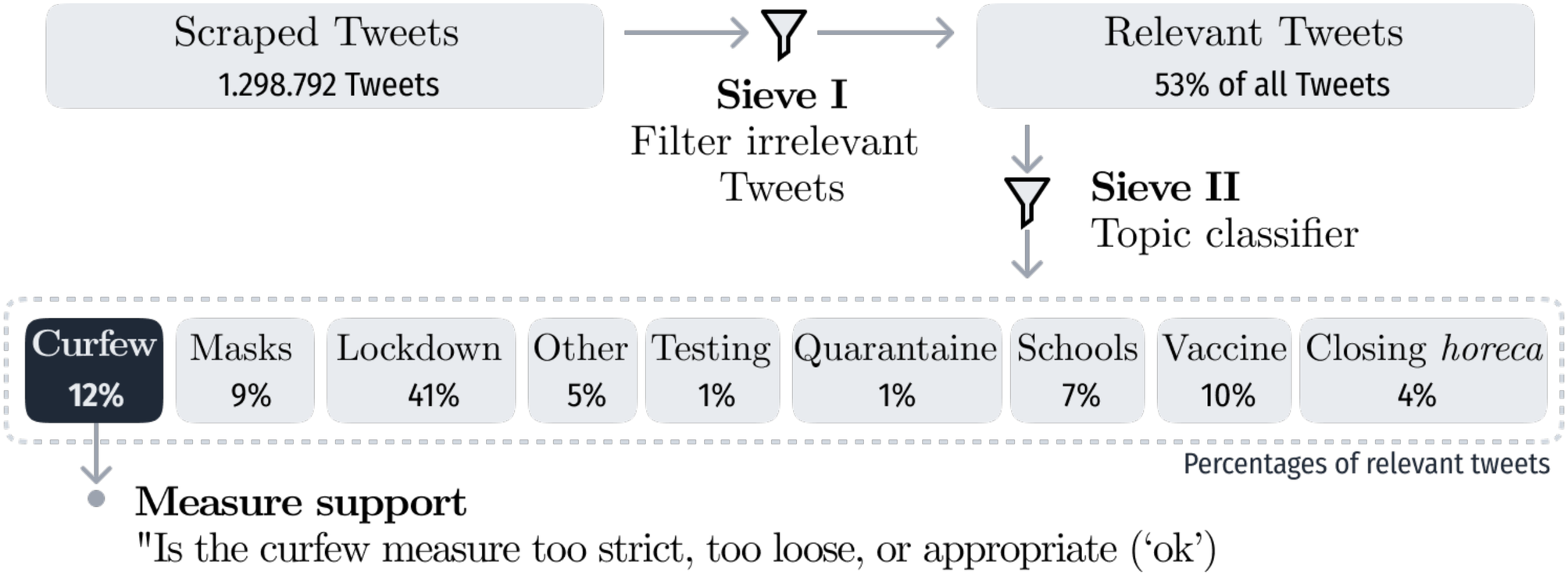
- Evaluation of 1.3M collected Tweets on COVID-19 measures
- Focused on discussion of COVID-19 policy in Belgium
- Additional focus on support for curfews
  - Belgium had multiple curfews (starting at midnight)

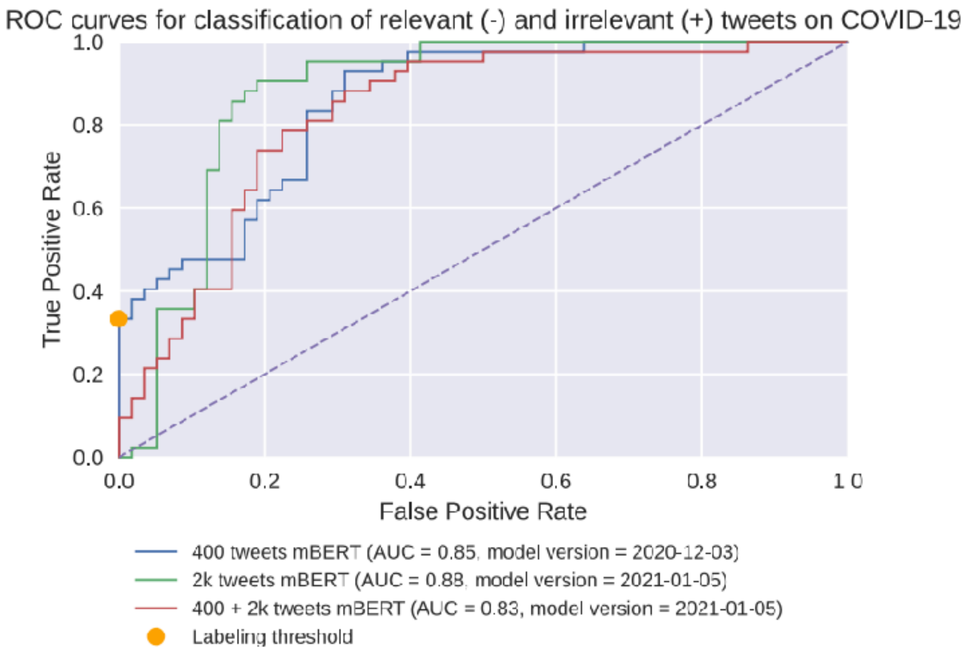
Kristen Scott, Pieter Delobelle, Bettina Berendt, 2021. *"Measuring Shifts in Attitudes Towards COVID-19 Measures in Belgium"*. Computational Linguistics in the Netherlands Journal

Labeling: Doccano

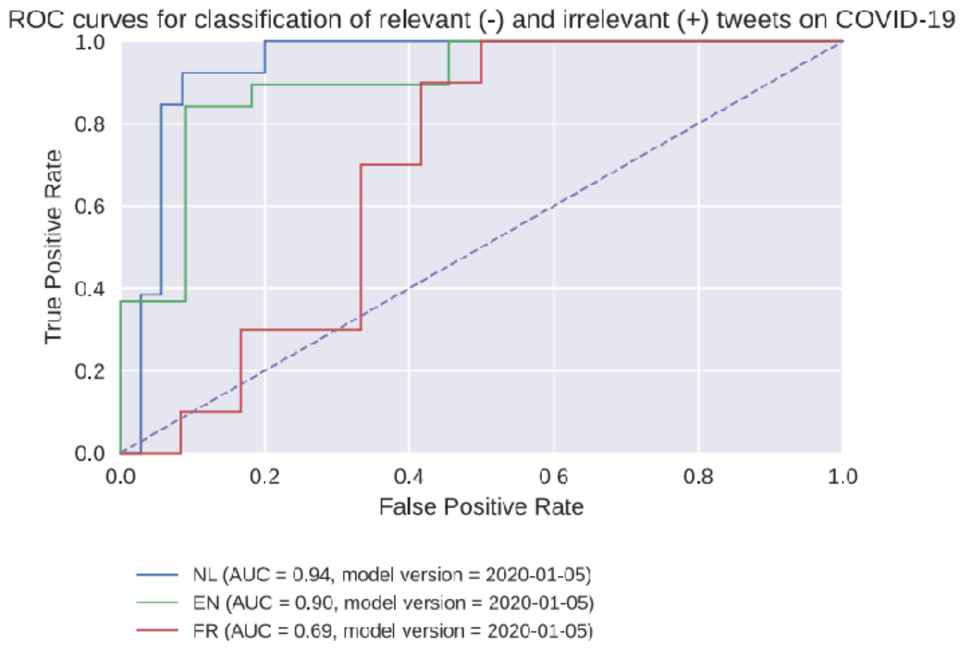


Our pipeline



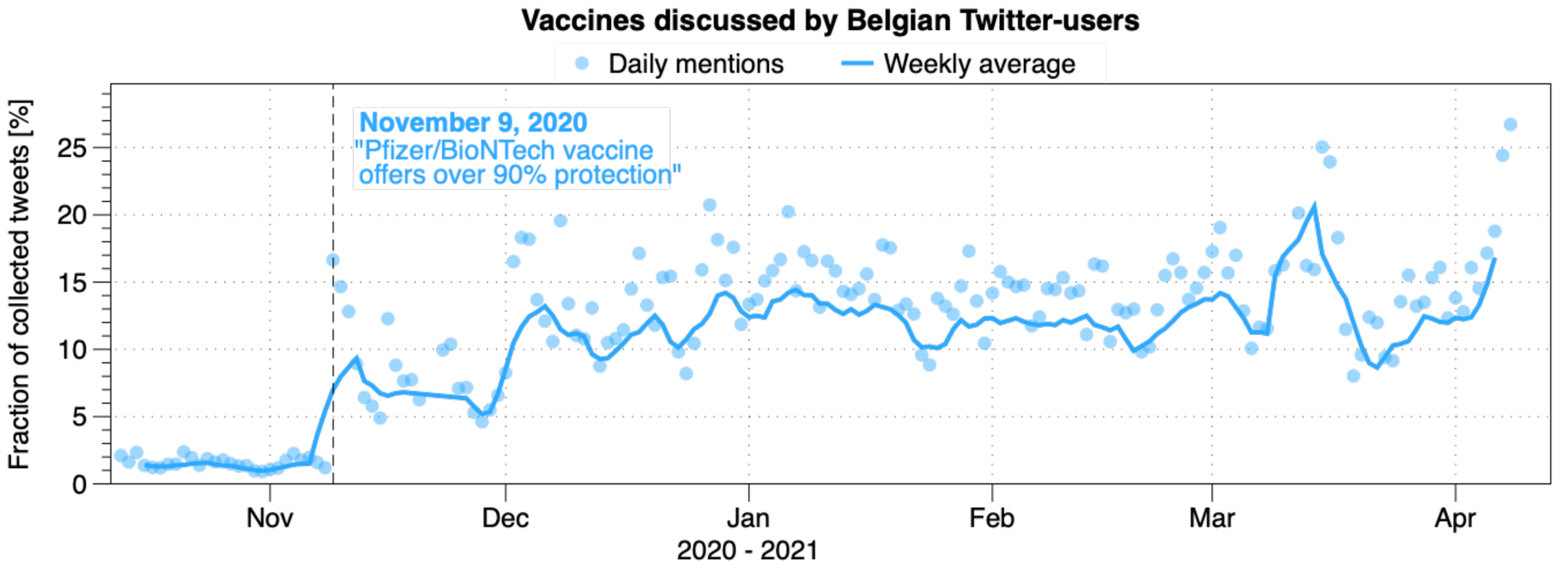


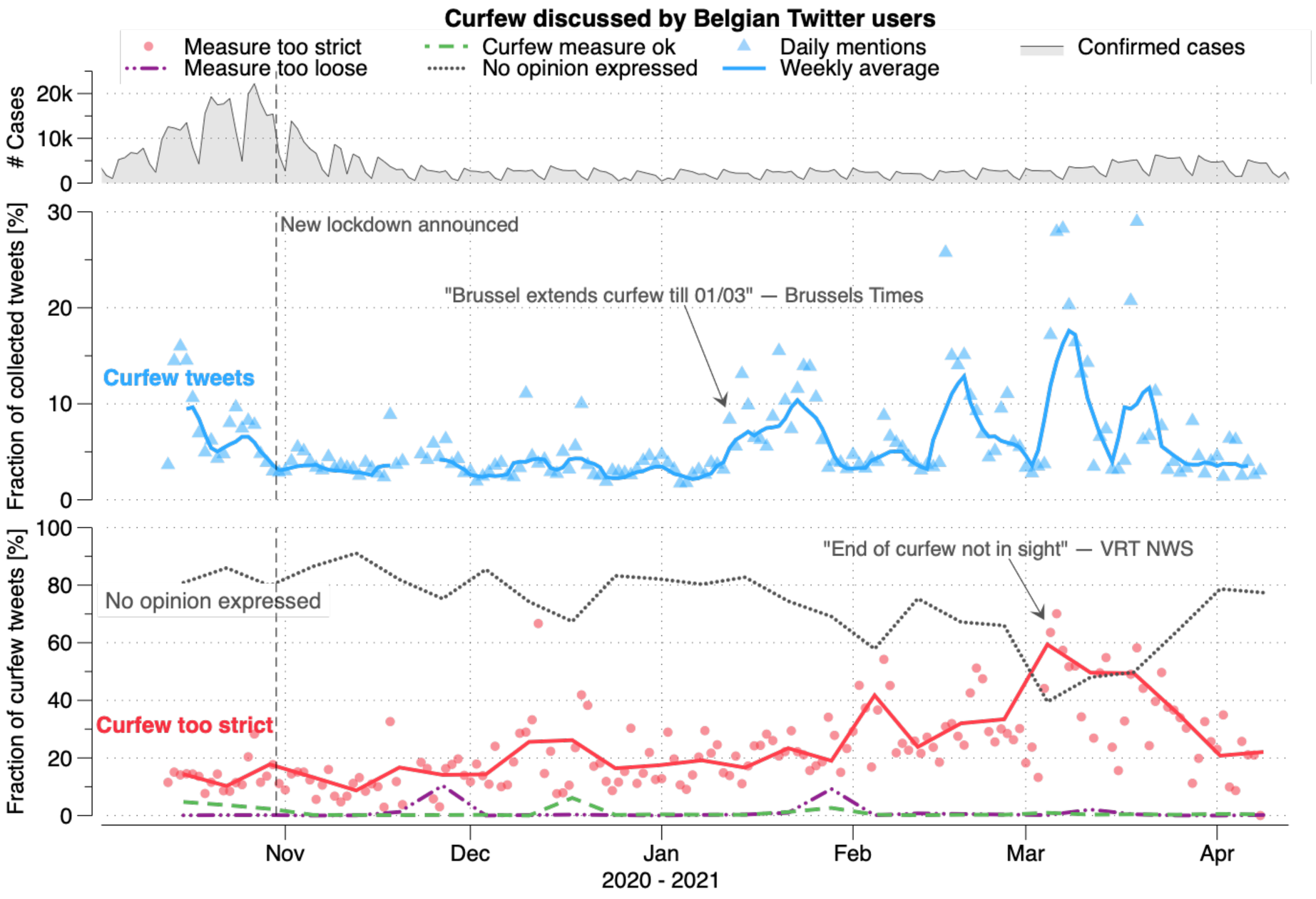
(a) ROC curves for different model versions, including the threshold set on the first (400 tweets) model used as Sieve 1.



(b) ROC curves conditioned on language (English, Dutch and French) for the best-performing model: mBERT trained on 2k tweets.

https://huggingface.co/DTAI-KULeuven/mbert-corona-tweets-belgium-topics


Vaccines discussed by Belgian Twitter-users
Daily mentions
Weekly average
November 9, 2020
"Pfizer/BioNTech vaccine
offers over 90% protection"
Fraction of collected tweets [%]
0
5
10
15
20
25
Nov
Dec
Jan
Feb
Mar
Apr
2020 - 2021


Curfew discussed by Belgian Twitter users
Measure too strict
Measure too loose
Curfew measure ok
No opinion expressed
Daily mentions
Weekly average
Confirmed cases
Cases
20k
10k
0
Fraction of collected tweets [%]
30
20
10
0
New lockdown announced
"Brussel extends curfew till 01/03" — Brussels Times
Curfew tweets
Fraction of curfew tweets [%]
100
80
60
40
20
0
"End of curfew not in sight" — VRT NWS
No opinion expressed
Curfew too strict
Nov
Dec
Jan
Feb
Mar
Apr
2020 - 2021

DTAI

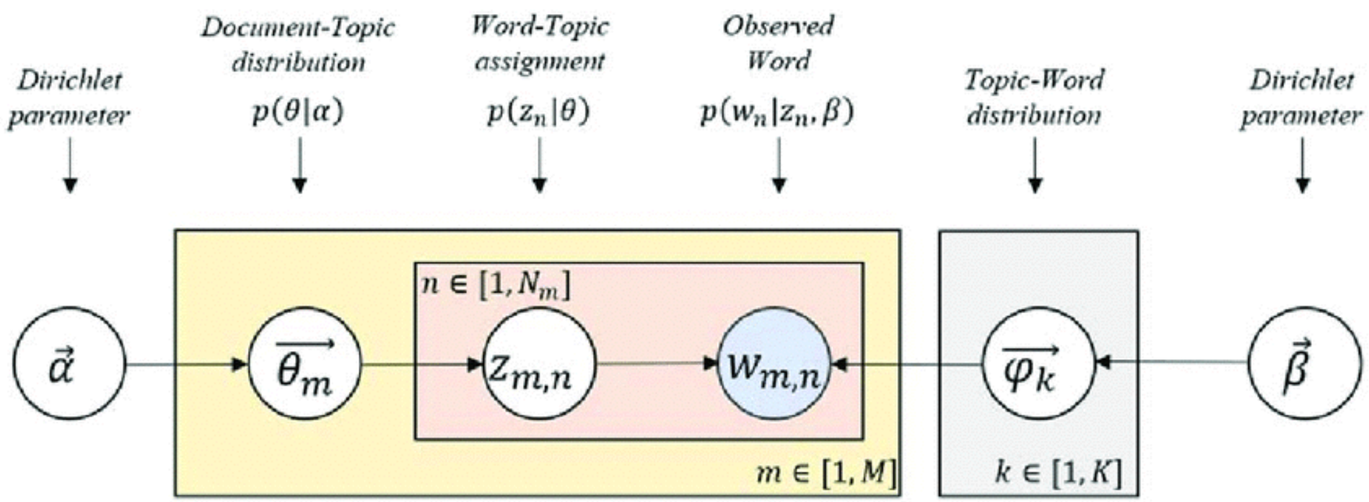
Topic modelling

Topic modelling

- Unsupervised machine learning
- Finding topics in a corpus
- Several approaches:
  - Latent Dirichlet Allocation (LDA)
  - BERTopic
  - …

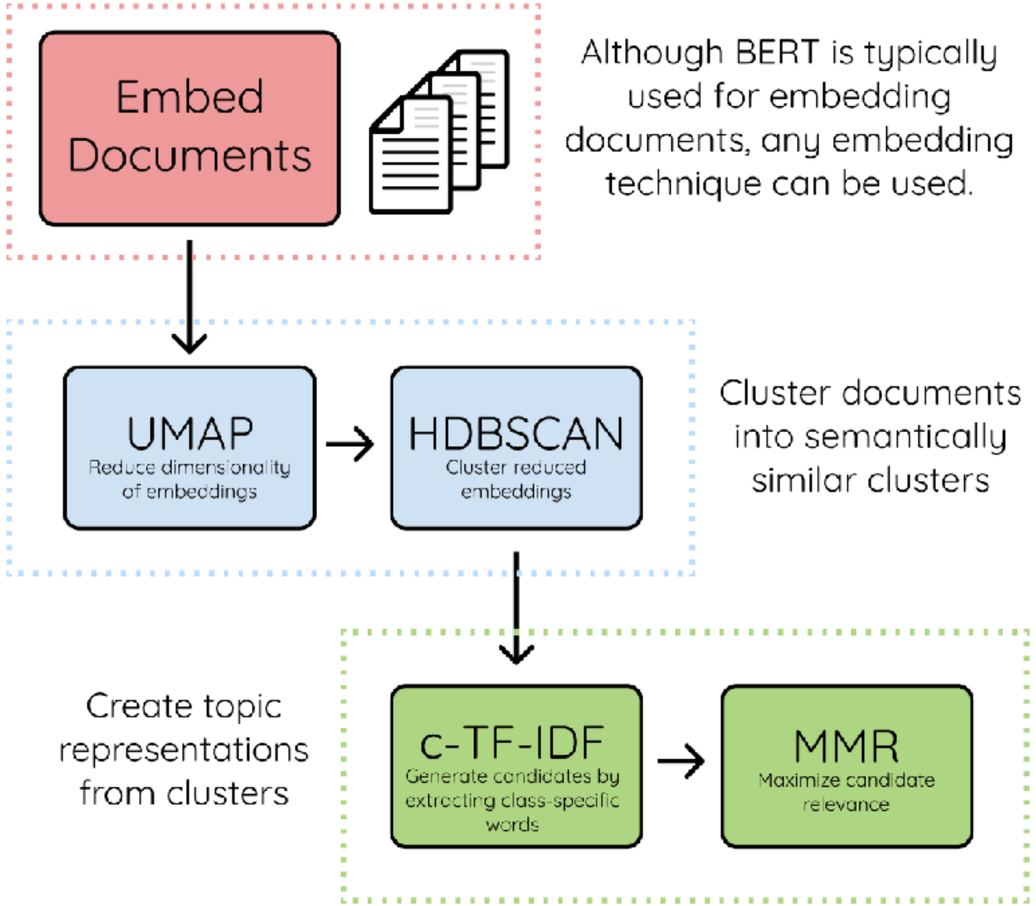
Latent Dirichlet Allocation

- Bayesian approach to model topics
  - words are generated by topics



BERTopic

- Topic modelling as clustering task
  1. Embedding with BERT
  2. Dim. reduction
  3. Clustering
  4. Get topic words



UMAP

- Dimensionality reduction
- Graph constructed from nearest neighbours

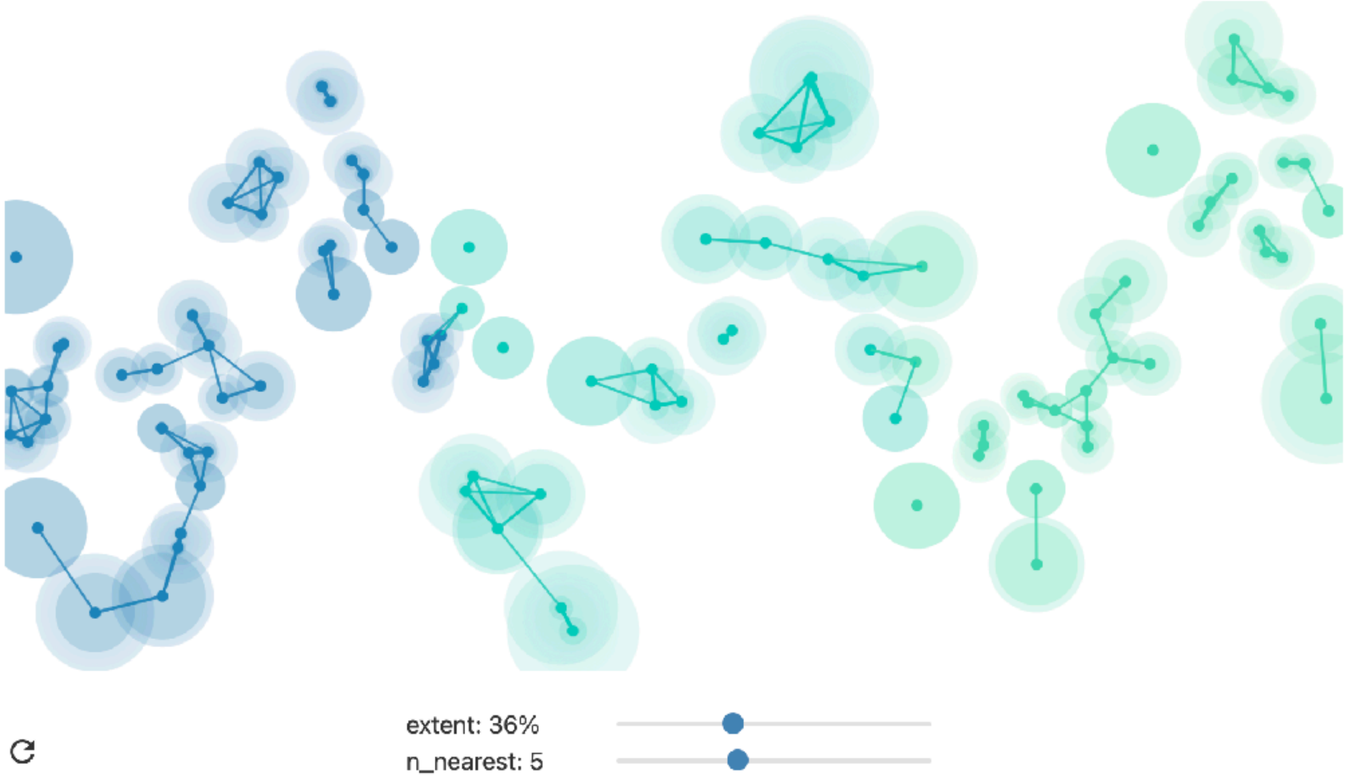


Figure 3: *Adjust the slider to extend a radius outwards from each point, computed by the distance to its nth nearest neighbor. Notice that past the intersection with the first neighbor, the radius begins to get fuzzy, with subsequent connections appearing with less weight;*

UMAP

- Dimensionality reduction
- Graph constructed from nearest neighbours
- Projected in 2D space by mapping the graph to 2D
  - High-dim. neighbours are closer together in 2D

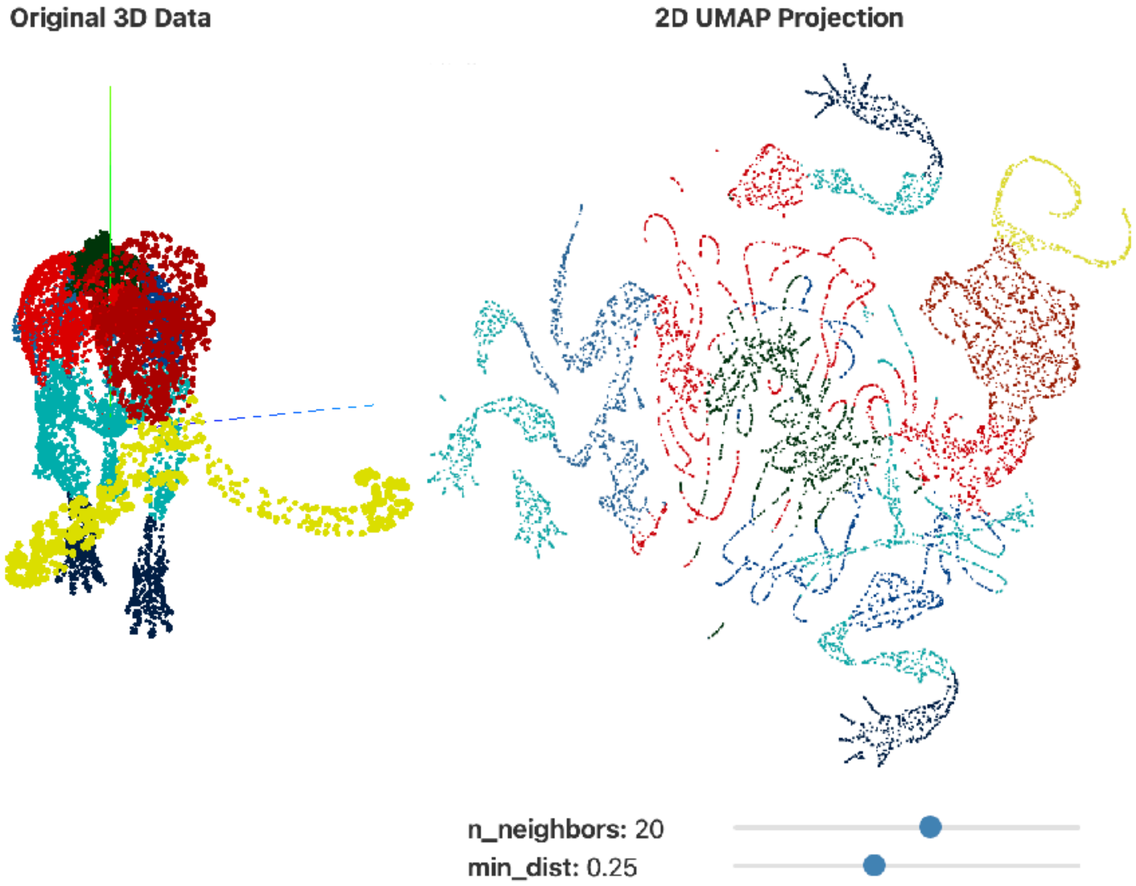


Figure 5: UMAP projections of a 3D woolly mammoth skeleton (50k points, 10k shown) into 2 dimensions, with various settings for the `n_neighbors` *and* `min_dist` *parameters.*

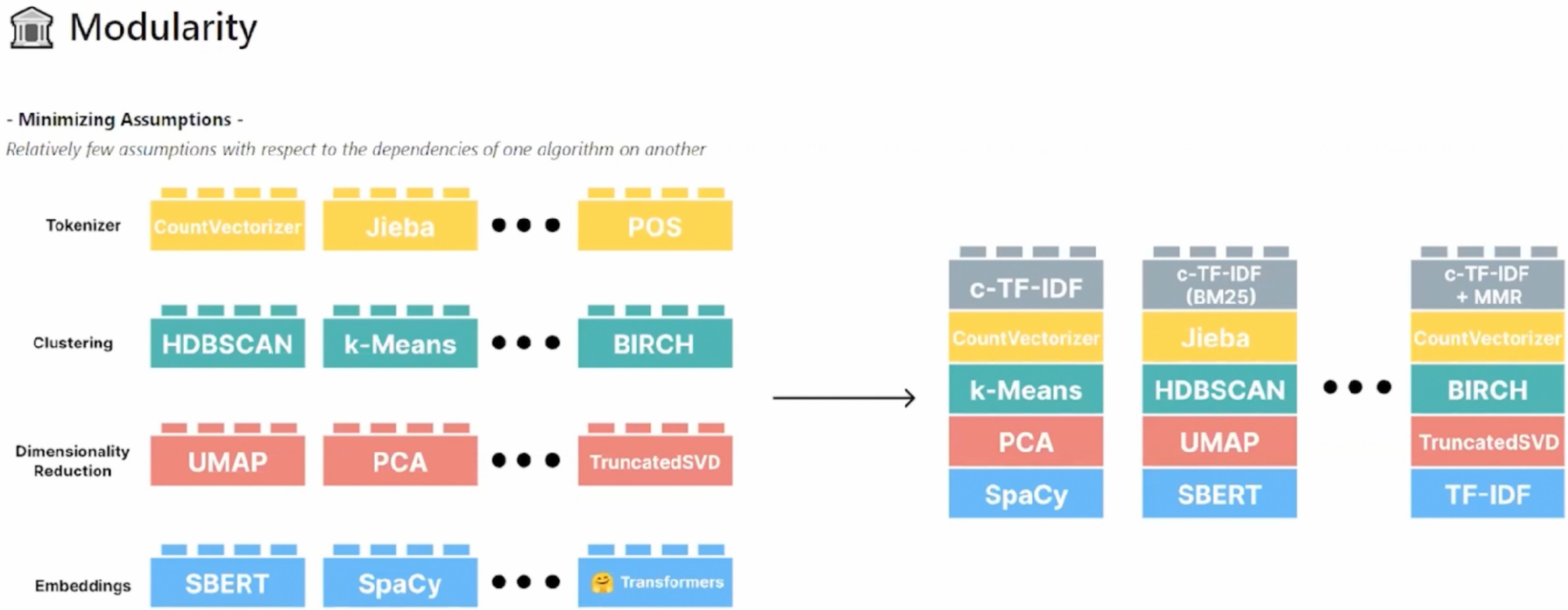
TF-IDF

- Measure of words that are unique for a document
- But corrected for words that are in every document (e.g. ‘the’)

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

BERTopic


Modularity
- Minimizing Assumptions -
Relatively few assumptions with respect to the dependencies of one algorithm on another
Tokenizer
CountVectorizer
Jieba
POS
Clustering
HDBSCAN
k-Means
BIRCH
Dimensionality Reduction
UMAP
PCA
TruncatedSVD
Embeddings
SBERT
SpaCy
Transformers
c-TF-IDF
CountVectorizer
k-Means
PCA
SpaCy
c-TF-IDF (BM25)
Jieba
HDBSCAN
UMAP
SBERT
c-TF-IDF + MMR
CountVectorizer
BIRCH
TruncatedSVD
TF-IDF