

# Deep Learning in Natural Language Processing for Analysis of Document Similarity

Pia Baronetzky, Noir Nigmatov, Theodor Stoican, and Peter Karl Weinberger

Technical University of Munich


Munich, 23th July 2021

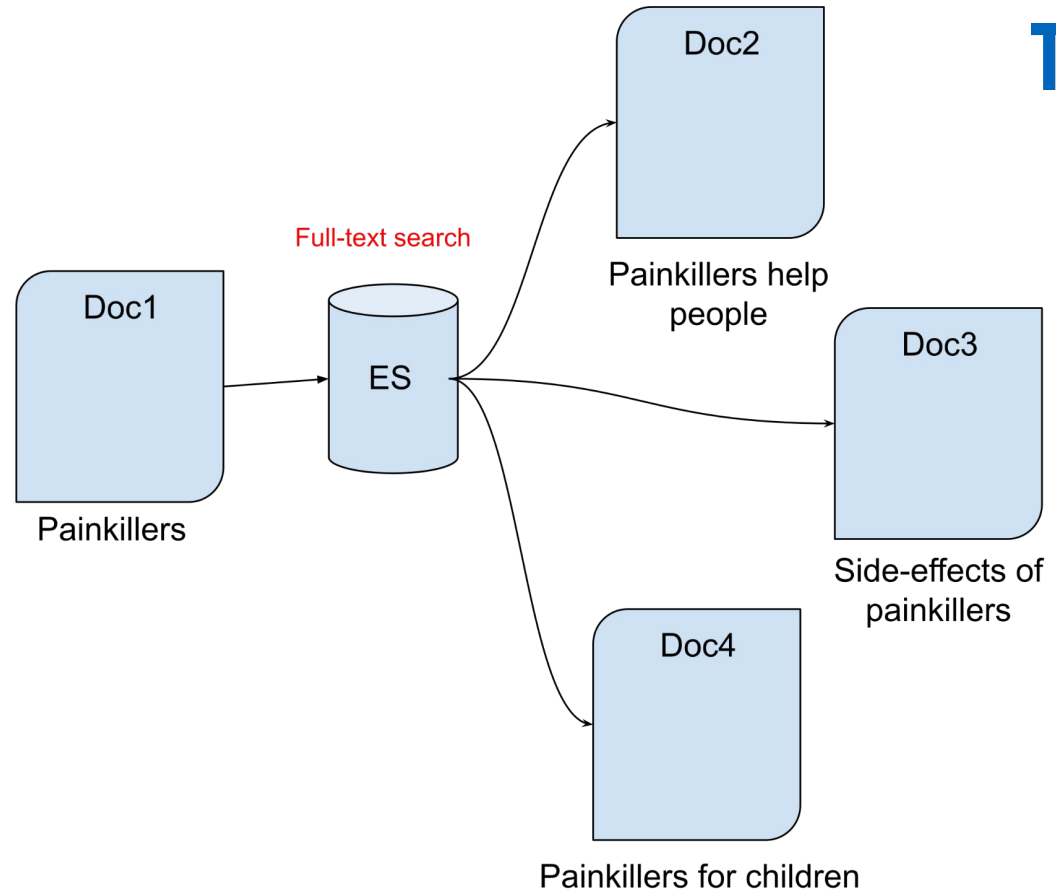


# Outline



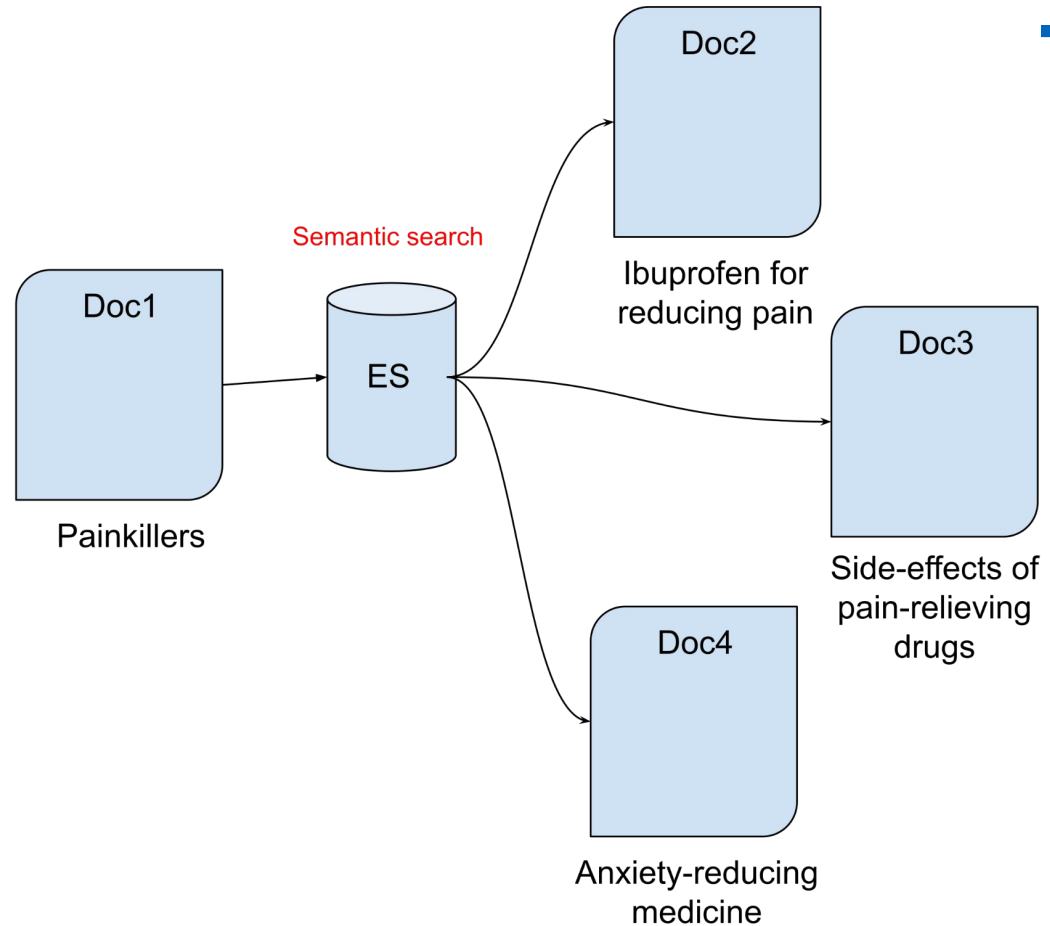
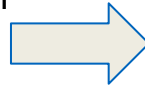
# Use-case

- Assume someone searches on Google "painkillers"
- Typically, they would get such a result as in the right 
- which typically includes "painkillers"



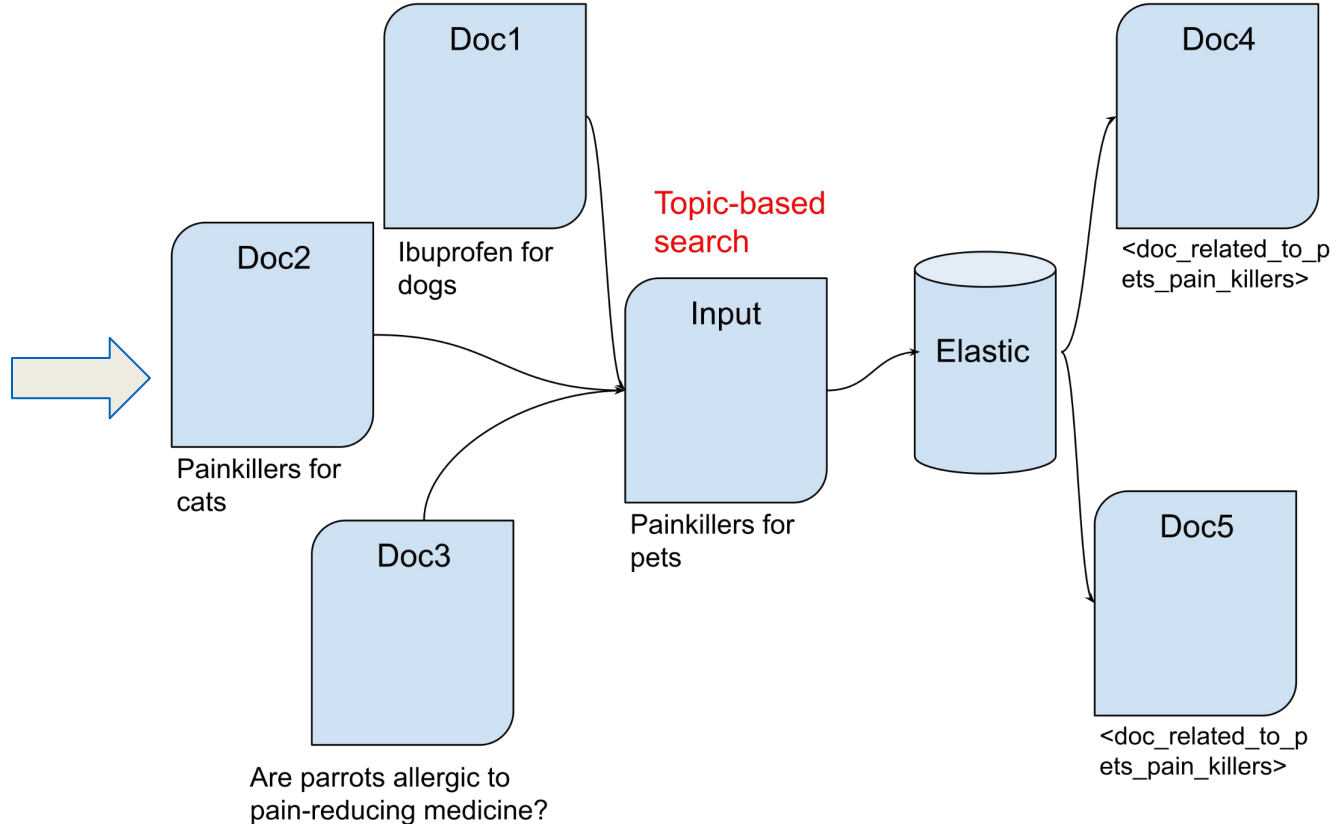
## Use-case(2)

- How could we make the results more semantically diverse?
- e.g. include instances of painkillers or semantically close words
- Answer: semantic search engine



## Use-case(3)

- Assume we also have multiple documents
- We want even more semantically general results
- Answer: Topic-based search



# In a nutshell – our contribution

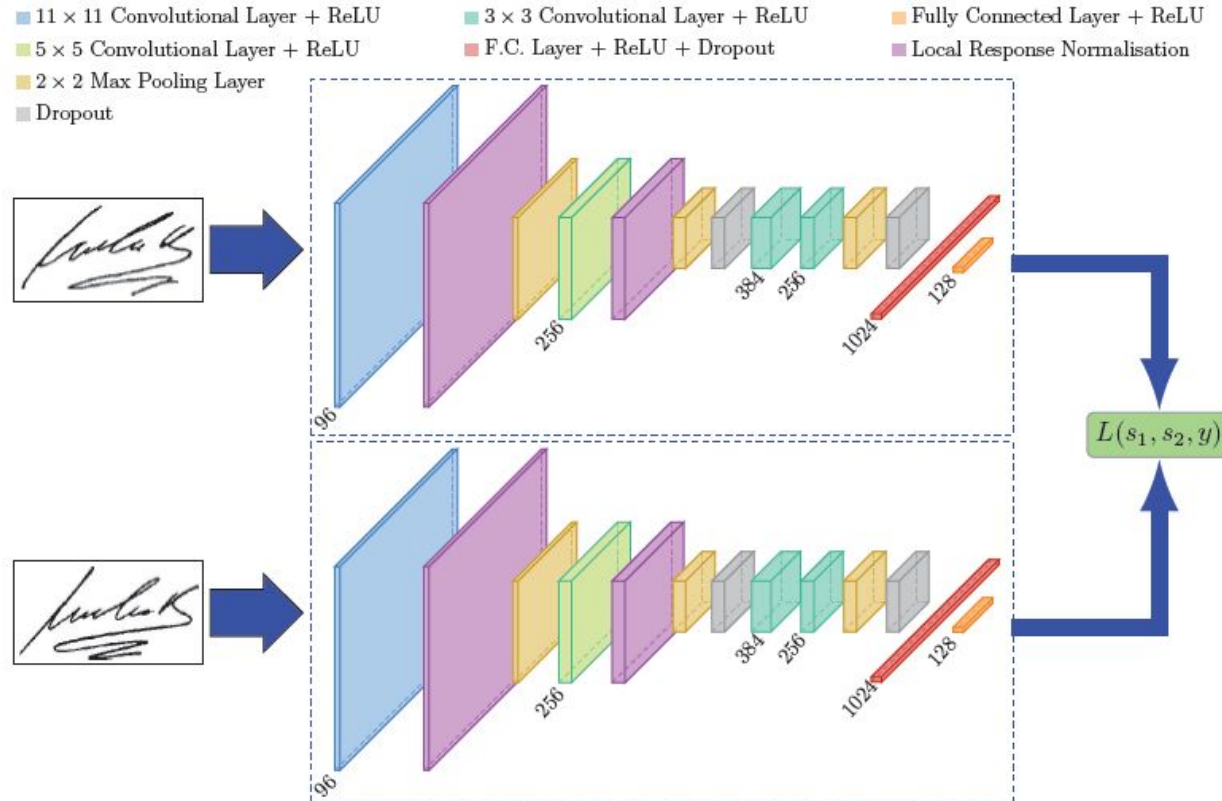
- Semantic search using latest insights from the NLP research
- Transformer-based topic-based search
- Combined weighted search
- Storage of data and fast retrieval with Elasticsearch
- Experimentation with re-ranking of the search results

# Sentence-BERT (Reimers et al., 2019)

- SOTA sentence embedding method
- SBERT: Sentence Embeddings using Siamese BERT-Networks
- SNLI + MNLI + STSb datasets
- Other previous embedding producing models:
  1. InferSent (Conneau et al., 2017)
  2. Universal Sentence Encoder (Cer et al., 2018)

# Siamese network

- uses the same weights while working in tandem on two different input vectors to compute **comparable output vectors**
- Similarity learning by the means of **Contrastive Loss**





















# Sentence-BERT

Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg.
Avg. GloVe embeddings	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
Avg. BERT embeddings	38.78	57.98	57.98	63.15	61.06	46.35	58.40	54.81
InferSent - GloVe	52.86	66.75	66.75	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder	64.49	67.80	67.80	76.83	73.18	74.92	76.69	71.22
SBERT-NLI-base	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT-NLI-large	72.27	78.48	74.90	80.99	76.25	79.23	73.75	76.55
SBERT-STSb-base	-	-	-	-	-	84.67	-	84.67
SBERT-STSb-large	-	-	-	-	-	84.45	-	84.45
SRoBERTa-NLI-base	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
SRoBERTa-NLI-large	74.53	77.00	73.18	81.85	76.82	79.10	74.29	76.68
SRoBERTa-STSb-base	-	-	-	-	-	84.92	-	84.92
SRoBERTa-STSb-large	-	-	-	-	-	85.02	-	85.02

Table 1: Spearman-rank correlation  $\rho$  between the cosine similarity of sentence representations and the gold labels for various semantic textual similarity (STS) tasks. Performance is reported by convention as  $\rho \times 100$ . STS12-STS16: SemEval 2012-2016, STSb: STSbenchmark, SICK-R: SICK relatedness data set.

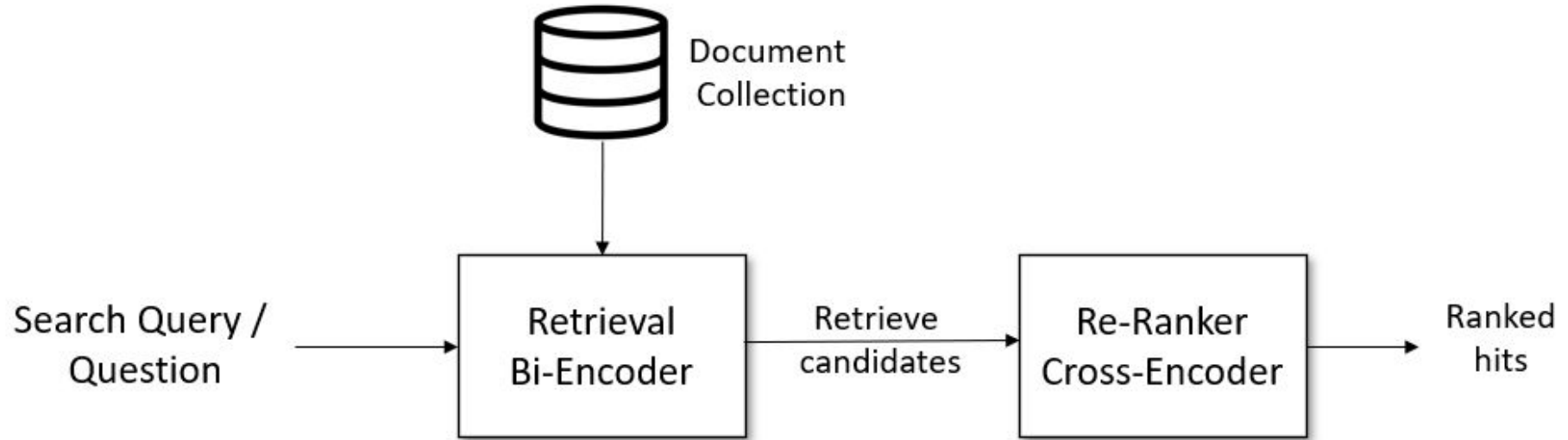
# Sentence-BERT pretrained models

Model Name	 STSb 	DupQ 	TwitterP 	SciDocs 	Clustering 	Avg. Performance 	Speed 
stsb-mpnet-base-v2 	88.57	85.04	75.35	72.48	39.16	72.12	2800
stsb-roberta-base-v2 	87.21	82.55	73.44	69.83	36.42	69.89	2300
paraphrase-mpnet-base-v2 	86.99	87.80	76.05	80.57	52.81	76.84	2800
paraphrase-multilingual-mpnet-base-v2 	86.82	87.50	76.52	78.66	47.46	75.39	2500
nli-mpnet-base-v2 	86.53	83.22	76.24	72.90	43.38	72.45	2800
stsb-distilroberta-base-v2 	86.41	82.70	73.68	69.85	37.68	70.07	4000
nli-roberta-base-v2 	85.54	80.20	74.28	69.86	40.12	70.00	2300
paraphrase-distilroberta-base-v2 	85.37	86.97	73.96	80.25	49.18	75.15	4000

# Experiments and selected models

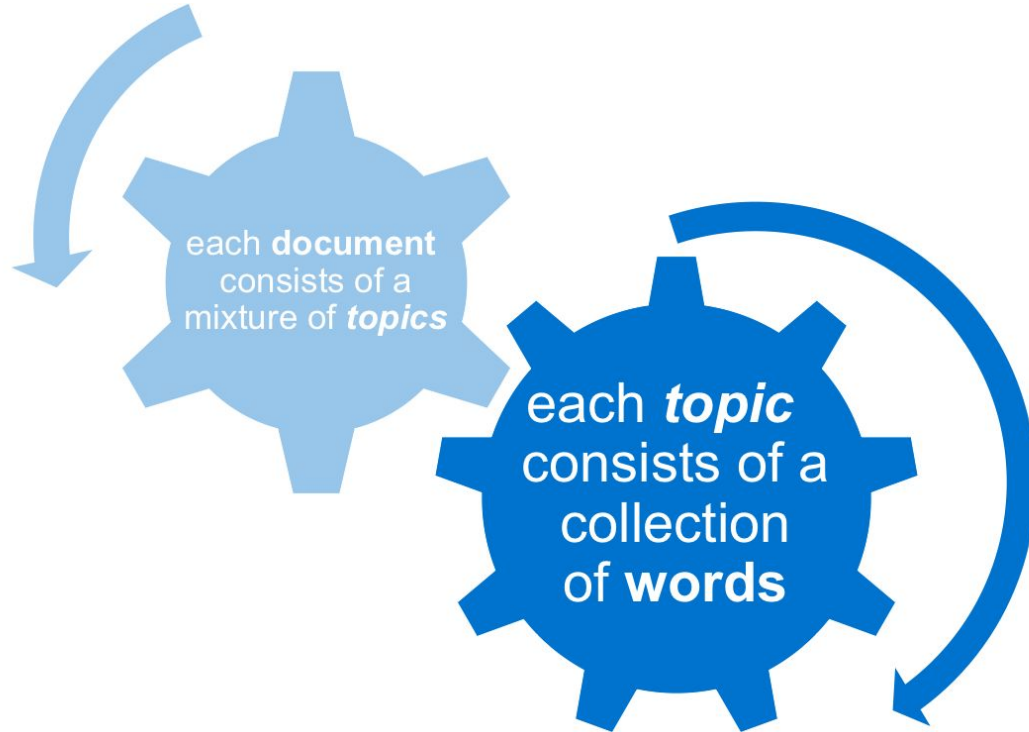
1. ***sts-b-mpnet-base-v2***: the base model is MPNET (Microsoft) [25]. Training data: NLI + STSb. It has the highest Spearman-rank correlation.
2. ***sts-b-distilroberta-base-v2***: the base model is DistilRoBERTa-base. Training data: NLI + STSb. The rank is 2 points less, but the inference speed is almost 1.5 times higher than the speed of mpnet.
3. ***distiluse-base-multilingual-cased-v1***: distilled version of the Multilingual Universal Sentence Encoder for 15 languages: Arabic, Chinese, Dutch, English, French, German, Italian, Korean, Polish, Portuguese, Russian, Spanish, Turkish. The rank is 80.62 and the inference speed is 4000 sentences/sec. This model was chosen because it is multilingual and the inference speed is the highest among other multilingual models though with slightly higher ranks.

# Retrieve and Re-rank



- selected Cross-encoder model: *cross-encoder/ms-marco-MiniLM-L-6-v2* (trained on MS MARCO)
- re-ranker is included as an option in our search engine

# Topic Models – A Tool for Semantic Search

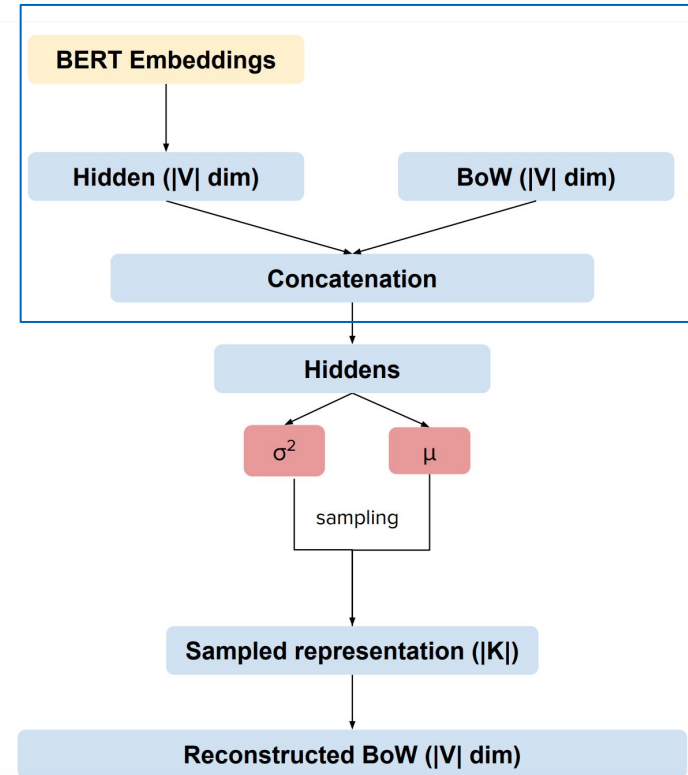


Both Models:

- Training on 100k documents
- Removal of stop words
- Final experiments with 10 and 50 topics

# Contextualized Topic Models

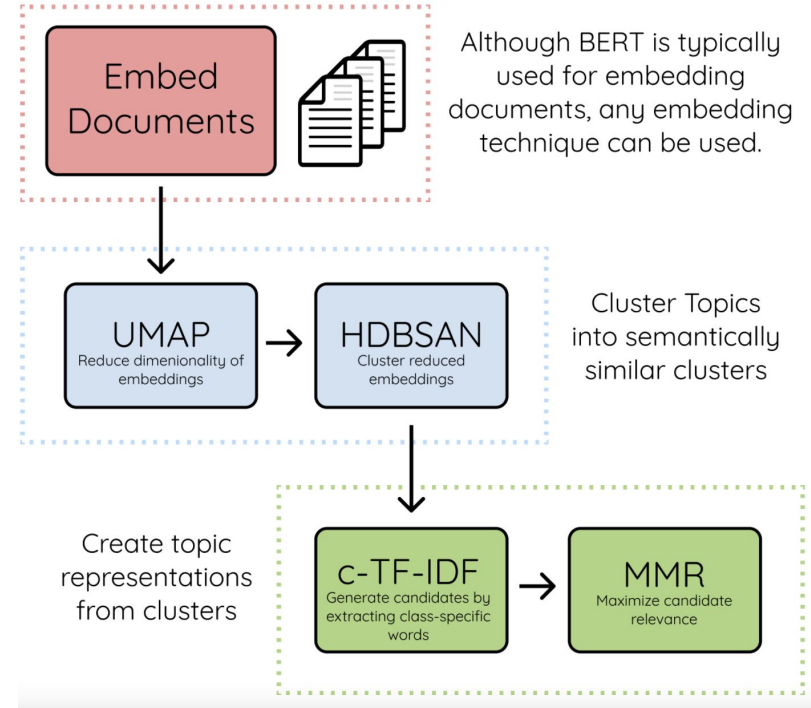
- Builds on ProdLDA
- Utilization of arbitrary sentence BERT embeddings
- ProdLDA
  - Variational autoencoder (VAE)
  - Neural inference network directly maps BoW representation into continuous latent representation
  - A decoder network reconstructs BoW by generating its words from the latent document representation
- Concatenation of BoW with context-aware sentence BERT embeddings as input
- Improvements to the code will be made publicly available



# BERTopic

In a nutshell:

- embed documents
- reduce their dimensionality (HDBSCAN is prone to the curse of dimensionality) - UMAP
- cluster them - HDBSCAN
- aggregate all documents from a cluster into a single one
- apply TF-IDF (this version is actually called c-TF-IDF).
- detect the most meaningful words within a cluster
- Voilà -> the topic



# Evaluation

- Literature-wise methods of evaluations are subjective
- Naked-eye comparison was performed in our case too
- CTM for 50 topics proved the most accurate

<b>Accuracy</b>	<b>CTM10</b>	<b>CTM50</b>	<b>BERTopic10</b>	<b>BERTopic50</b>
	70%	90%	40%	70%

Table 2: Naked-eye comparison of the 4 topic models.



# Demo

# Conclusion

