

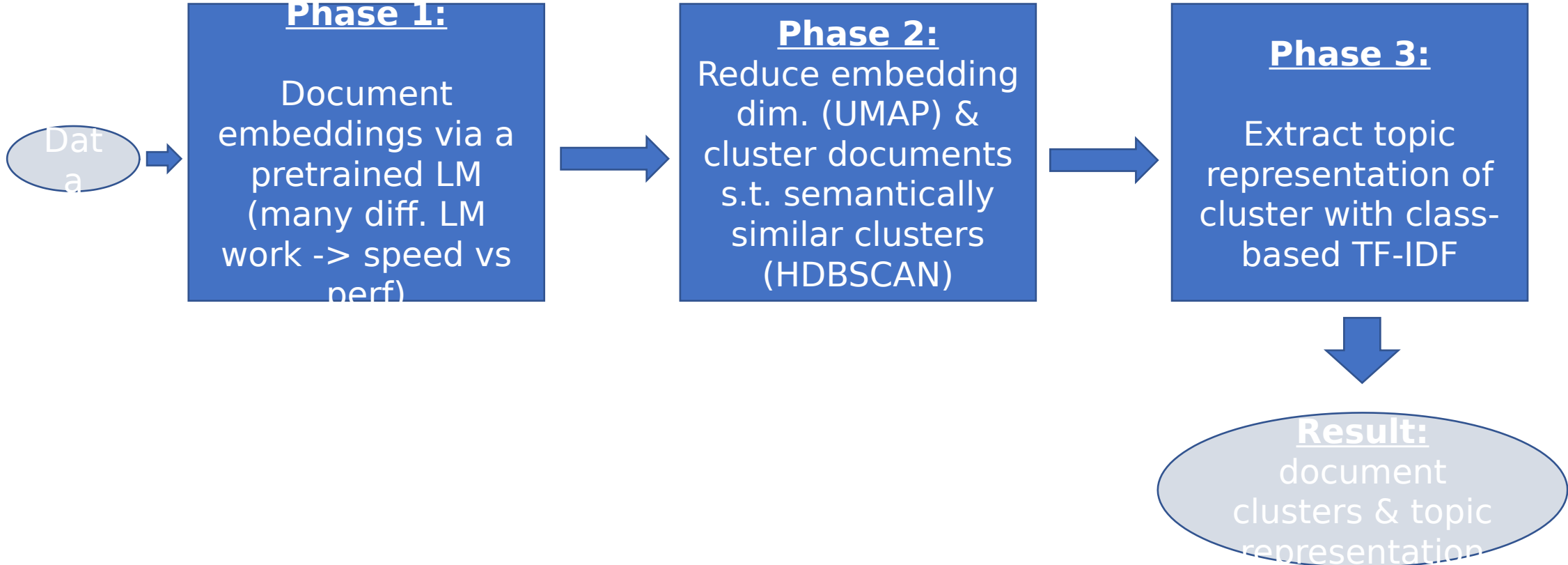
BERTopic

a short introduction

Motivation

- Compared to classical methods (LDA, NMF) that represent docs as „bag-of-words“: embedding-based clustering approach
- Difference to previous embedding-based approaches:
How to represent cluster?
Words close to cluster centroid VS *BERTopic*: modified TF-IDF

Algorithm



Details #1

- Used LM: Sentence-BERT framework (SBERT) -> algorithm „scales“ with future/better LMs
- Used cluster alg: HDBSCAN – hierarchical soft clustering approach
 - > allows unrelated docs to be assigned to no cluster (modelled as noise in the data)
 - > empirically shown: UMAP & HDBSCAN work well together

Details #2

Classic TF-IDF: for term (word) t and doc d

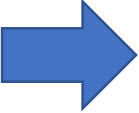
Term-frequency,
e.g. $\frac{\#\{t \text{ in } d\}}{\#\{\text{words in } d\}}$

Inverse doc
freq.:
 $\frac{\#\text{docs}}{\#\{\text{docs with } t\}}$

Class-based TF-IDF: concat all docs in a cluster to single doc \rightarrow importance of term t in cluster c :

Avg
#words
per doc

Frequency
of term t in
all clusters

 Generate topic-word distributions for each cluster of documents;
Can achieve $\# \text{topics} = \text{preset value}$ BY merging least common topic with most similar topic (based on c-TF-IDF)

Experimental Setup

- Uses OCTIS for comparison of algs (data, preprocessing, results)
- Datasets: 20NewsGroups, BBCNew and Trump's tweets
- Preprocess for first two: remove punctuation/stopwords/docs with #words < 5, lemmatization, lowercase all tokens
- Compare BERTopic with different LMs and LDA, NMF, CTM, Top2Vec
- **Evaluation:** *topic coherence* with NPMI [-1,1] and *topic diversity* with % of unique words for all topics [0,1]
<-> are only an indicator of human judgement & running times

Results

	20 NewsGroups		BBC News		Trump	
	TC	TD	TC	TD	TC	TD
LDA	.058	.749	.014	.577	-.011	.502
NMF	.089	.663	.012	.549	.009	.379
T2V-MPNET	.068	.718	-.027	.540	-.213	.698
T2V-Doc2Vec	.192	.823	.171	.792	-.169	.658
CTM	.096	.886	.094	.819	.009	.855
BERTopic-MPNET	.166	.851	.167	.794	.066	.663

Table 1: Ranging from 10 to 50 topics with steps of 10, topic coherence (TC) and topic diversity (TD) were calculated at each step for each topic model. All results were averaged across 3 runs for each step. Thus, each score is the average of 15 separate runs.

	20 NewsGroups		BBC News		Trump	
	TC	TD	TC	TD	TC	TD
BERTopic-USE	.149	.858	.158	.764	.051	.684
BERTopic-Doc2Vec	.173	.871	.168	.819	-.088	.536
BERTopic-MiniLM	.159	.833	.170	.802	.060	.660
BERTopic-MPNET	.166	.851	.167	.792	.066	.663

Table 2: Using four different language models in BERTopic, coherence score (TC) and topic diversity (TD) were calculated ranging from 10 to 50 topics with steps of 10. All results were averaged across 3 runs for each step. Thus, each score is the average of 15 separate runs.

Results #2

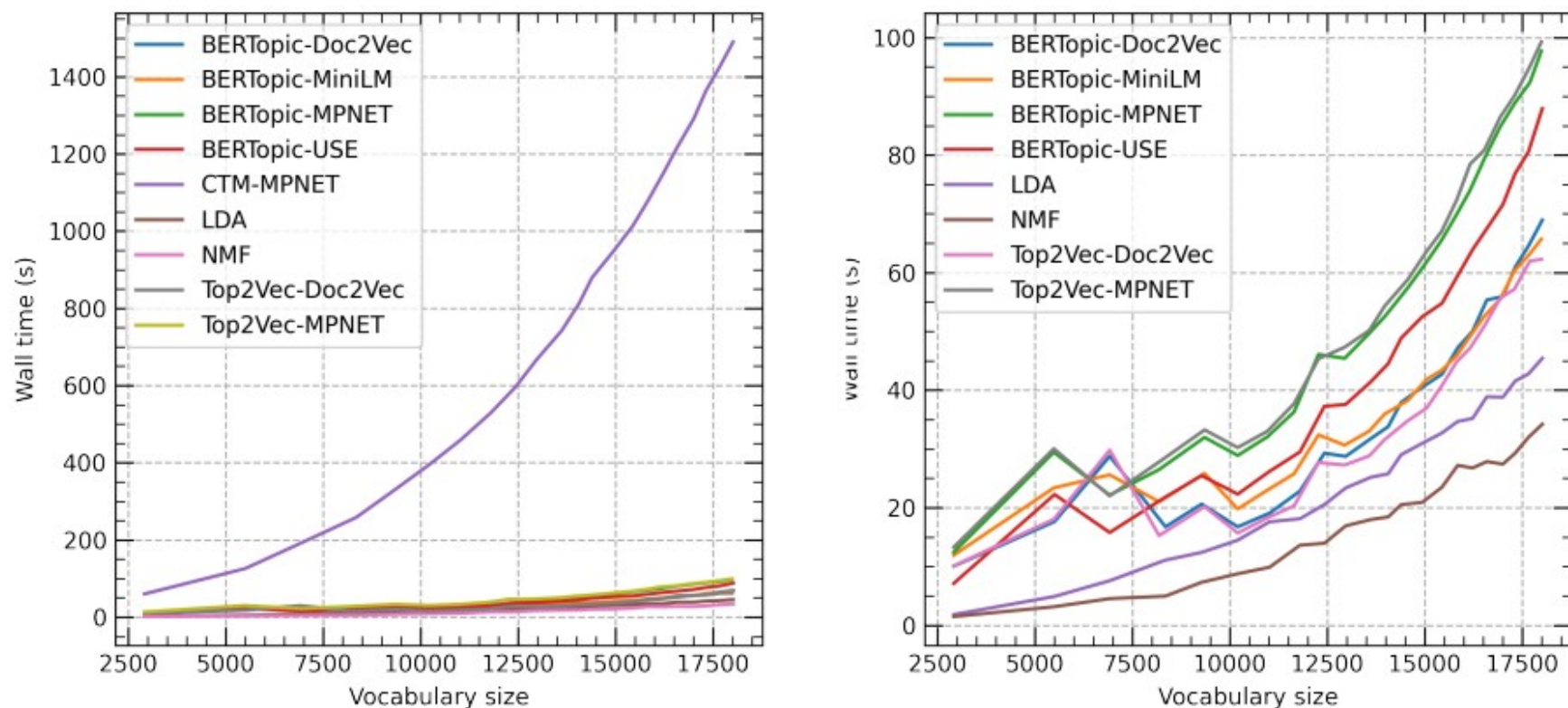


Figure 1: Computation time (wall time) in seconds of each topic model on the Trump dataset. Increasing sizes of vocabularies were regulated through selection of documents ranging from 1000 documents until 43000 documents with steps of 2000. **Left:** computational results with CTM. **Right:** computational results without CTM as it inflates the y-axis making differentiation between other topic models difficult to visualize.

Strengths/Weaknesses

+	-
<ul style="list-style-type: none">• Competitive with all LMs used here• Separation of embedding docs and representing topics• Represent topics as distribution of words (via c-TF-IDF)	<ul style="list-style-type: none">• Assumes one topic per document• Topic representation is with „bag-of-words“ approach VS contextual representation -> makes words describing a topic often similar

References

**BERTopic: Neural topic modeling with a class-based
TF-IDF procedure**

<https://arxiv.org/abs/2203.05794>