Comparative Analysis of Generative and Neural Topic Modeling Techniques Applied to Modern Greek Political Corpora

Nikolaos Giannakopoulos | Dionysios Rigatos

TDT4310

April 2024

Norwegian University of Science and Technology

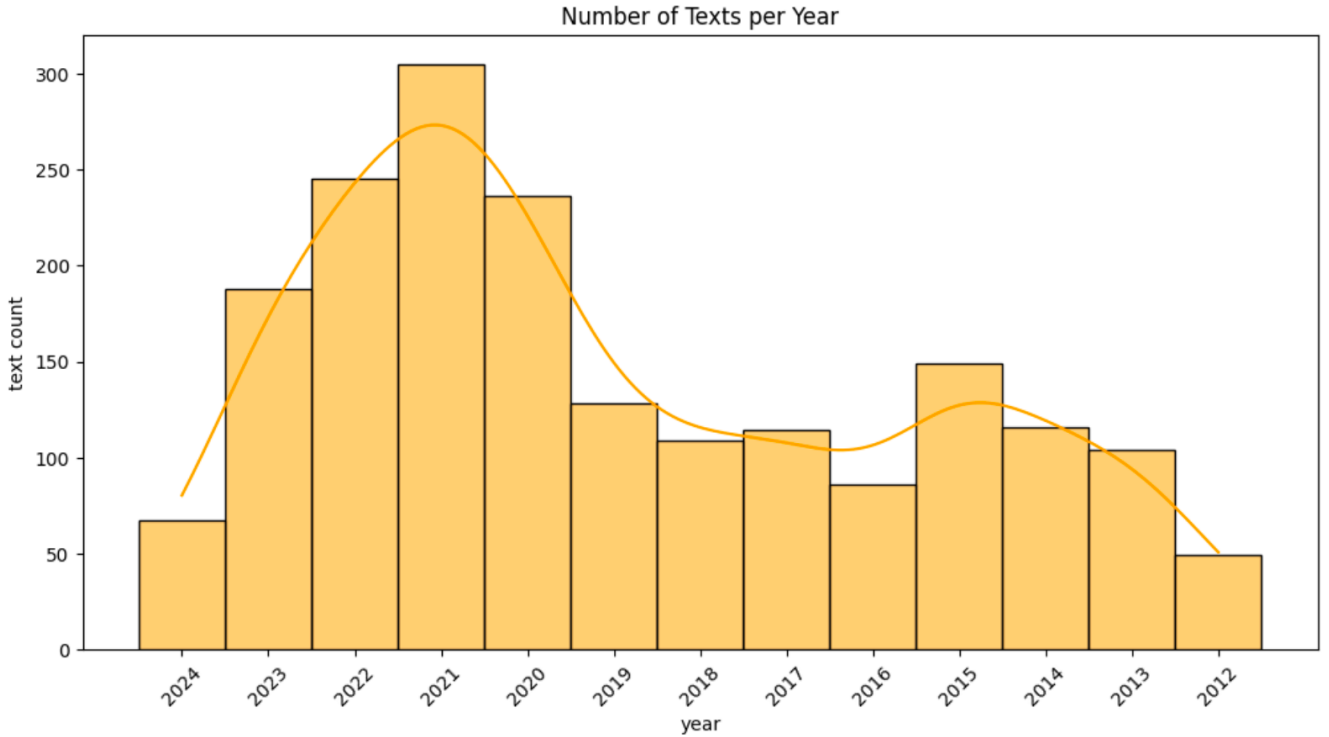
Motivation & Goals

- Topic modeling is not a buzzword but has a lot of depth and interesting applications
- Multilingual tasks are more accessible than ever
  - We are Greek!
- Initially an analytical task, quickly diverged into an optimization task
  - Can classic approaches compare with SOTA?
  - Can topic models be quantitatively evaluated?
  - How to achieve the best possible qualitative results?
- Unsupervised approach

Dataset

- Statements and speeches from the Office of the Greek Prime Minister in the period 2012-2024
- Primarily in Greek, minor deviations
- No labels or splits
- Manually scraped
- Politics are complicated; we generate 30 topics per algorithm
- ~2000 documents, usually quite long

Dataset



NTNU

Approach I - OCTIS

- **O**ptimizing and **C**omparing **T**opic models **I**s **S**imple: provides pre-processing, training and evaluation for most common topic modelling algorithms
- Many algorithms;
  - Experimented with Latent Semantic Indexing (LSI), Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), Non-Factorial Matrix Factorization (NMF) and more
  - Focus on **Product of Experts LDA (ProdLDA)**

OCTIS - Data Preprocessing

- Stopword, punctuation and number removal
- Minimum word length of 4
- TF-IDF weighting – max frequency of 20%, min of 1%
- Lemmatization
- Only applied to OCTIS models; BERTopic does not require preprocessing

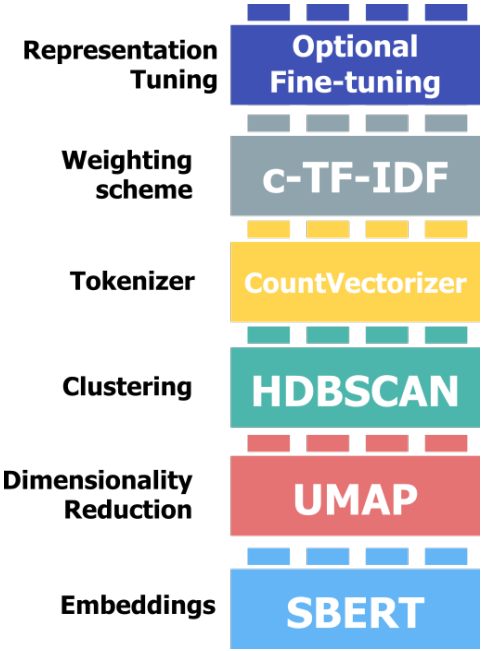
OCTIS - Issues

- **Problem:** OCTIS' spaCy-based preprocessing did not work well, had to find alternative
- **Solution:** Custom pipeline with Stanza and custom POS-based lemmatizer tailored for Greek data
- Adapted for compatability with the library
- Impressive results

Approach II - BERTopic

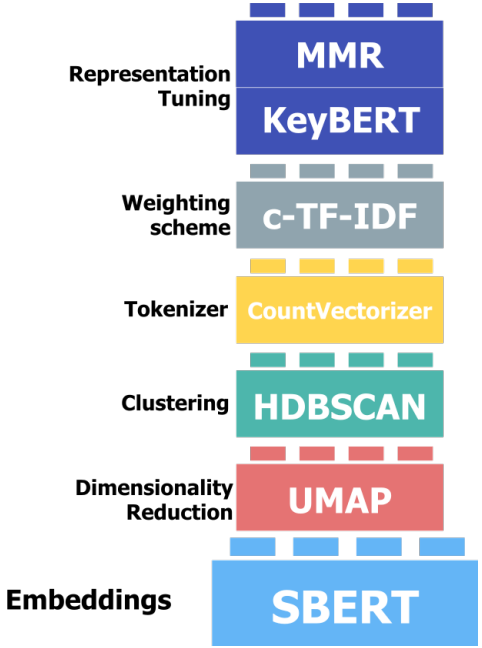
- **BERTopic**: Neural topic modelling, modular framework based on Sentence-BERT

BERTopic - Architecture



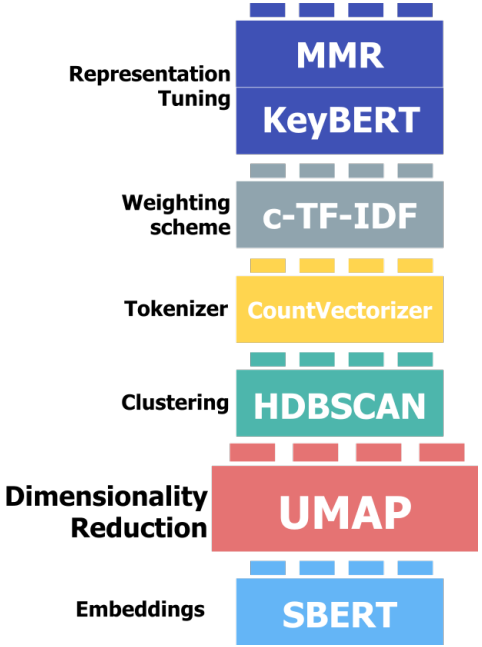
- **BERTopic**: Neural topic modelling, modular framework based on Sentence-BERT
- Can be fine-tuned at any stage
- Results mainly rely on clustering & weighting scheme
- Hyperparameter optimization for clustering and dimensionality reduction

BERTopic – Sentence Transformers



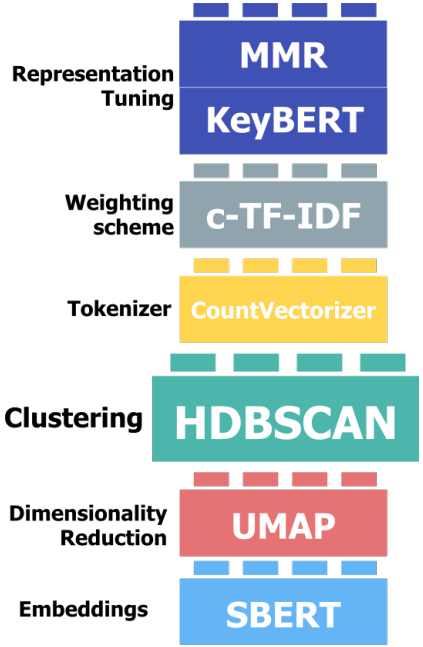
- **Sentence-BERT**: Pre-trained transformer embeddings
- Turns entire sentences/documents to embeddings
- Different models available online
- Opted for a model trained on Greek media

BERTopic – Dimensionality Reduction



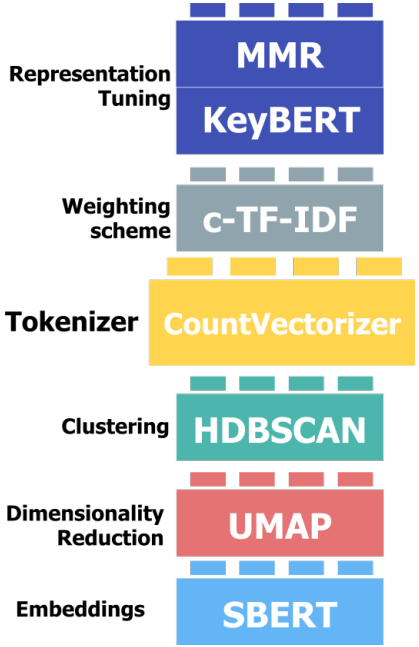
- Embeddings are high-dimensional; dimensionality reduction necessary
- Solution: **UMAP**
  - Non-linear dimensionality reduction, unlike PCA
  - Preserves local features better than other popular techniques (PCA, t-SNE etc.)
  - Confirmed by experiments and optimization

BERTopic – Clustering



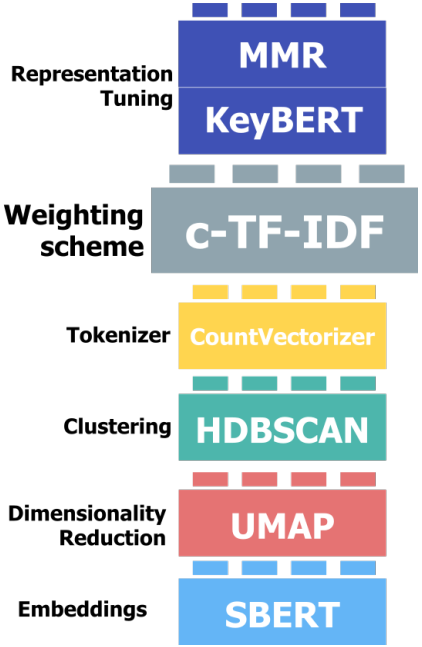
- Clusters define topics in BERTopic
- Solution: **HDBSCAN**
  - Covers a variety of distributions
  - Identifies dense regions of points within a radius from each other
  - Search for varying radius values; clusters of varying densities are found
  - Performs well on lower dimensions

BERTopic – Tokenizer



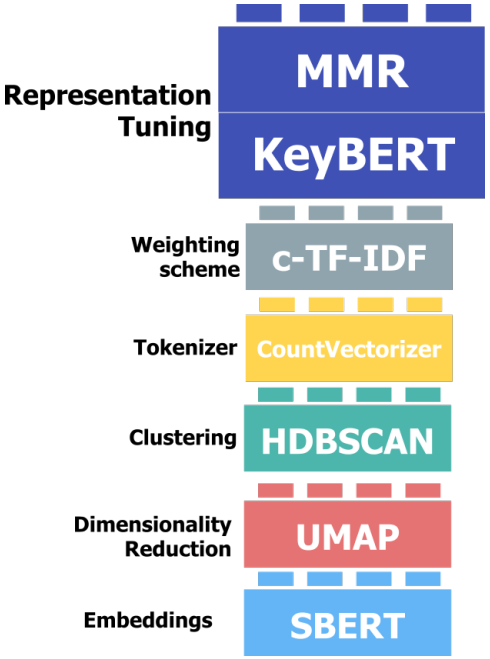
- Clusters still contain noise and trivial words
- Typical processing, applied post-clustering
  - Stopword removal, cleaning based on document frequency, (1-2)-grams
  - Unlike OCTIS, soft filtering on document frequency – max document frequency of 95% and min of 0.5%

BERTopic – Weighting Scheme



- Each cluster may include hundreds of words, need to find most representative
- **Class based TF-IDF**, word frequency on cluster level instead of document
- Last mandatory pipeline stage, final topics are produced here

BERTopic – Representation Tuning



- c-TF-IDF relies on word frequency → Semantically relevant words may be excluded
- **KeyBERT**: Finds most representative words in a cluster
- **Maximal Marginal Relevance**: Measures both relevance and diversity of words within a cluster
- Optional step in the pipeline

Hyperparameter Optimization

- Plethora of models and combinations, impossible to find optimal
- Hyperparameter optimization for all models in pre-defined search spaces
  - 5 hours for OCTIS models
  - 10 hours for BERTopic
- We optimized based on **CV Coherence**
- TL;DR: Classic models require a lot of tuning, BERTopic's defaults are acceptable

Evaluation Metrics

- Topic modelling is hard to evaluate quantitatively, lots of metrics exist
- **Coherence**: How well topics can be explained by the corpora they were generated from
- **Diversity**: Lexical diversity across topics
- **Similarity**: How similar topics are between one another
- **Human evaluation** is necessary

Evaluation

Table 1: Quantitative Topic Evaluation Metrics for All Models

Model	Coherence_{CV}	Coherence_{UMASS}	Diversity_{TOPIC}	Similarity_{PJS}
LSI	0.5585	-1.3672	0.5714	0.0334
LDA	0.6418	-1.3635	0.7762	0.0132
HDP	0.4650	-2.2493	0.4905	0.0162
NMF	0.5703	-1.6358	0.5143	0.0350
ProdLDA	**0.6782**	-2.8220	0.8857	0.0069
BERTopic	0.5996	**-0.5838**	0.9523	0.0020
BERTopic+	0.2903	-0.6087	**0.9714**	**0.0012**

BERTopic+ refers to the BERTopic model with additional Topic Representation Tuning.

Evaluation

Topic	ProdLDA	BERTopic+	BERTopic
1	Obligated, clearance, entrance, minute, overcharge, mouth, enumerate, analyst, objectively	Government, solution, state, reforms, debt, you know, national	*democracy, reforms*, euro, region, development, European, agreement
2	Energy, presidency, directive, regulation, ultimate, ready, multi-year, adoption, migration, unemployment	Stability, memorandums, government, parliament, reforms, debt, Greek people	Tsipras, elections, Mitsotakis, SYRIZA, euro, Mr. Mitsotakis, ladies gentlemen
3	Partner, negotiation, regulation, radical, administration, suffocating, institutional, eurozone, claiming, tax	Health system, doctors, nurses, beds, ICU, hospitals, health, Vasilis, vaccination, intensive care	Health, hospital, health system, hospitals, vaccination, national system, system
4	Freedom, democratic, tolerance, against, racism, order, stereotypes, extremism, enemy, ethos	Prime Minister, Antonis, democracy, privileged, national, interests, simply, EAB, presidency	*que, de, en, la, el*, Samaras, democracies
5	Clearance, obligated, disastrous, mouth, entrance, writing, stock market, similar, apportion	Minimum wage, primary, new, CAP, minimum, livestock farmers, restaurant industry, businesses	Farmers, agricultural, products, agricultural, primary, sector, products, disability

Evaluation

Topic	ProdLDA	BERTopic+	BERTopic
1	Obligated, clearance, entrance, minute, overcharge, mouth, enumerate, analyst, objectively	Government, solution, state, reforms, debt, you know, national	*democracy, reforms*, euro, region, development, European, agreement
2	Energy, presidency, directive, regulation, ultimate, ready, multi-year, adoption, migration, unemployment	Stability, memorandums, government, parliament, reforms, debt, Greek people	Tsipras, elections, Mitsotakis, SYRIZA, euro, Mr. Mitsotakis, ladies gentlemen
3	Partner, negotiation, regulation, radical, administration, suffocating, institutional, eurozone, claiming, tax	Health system, doctors, nurses, beds, ICU, hospitals, health, Vasilis, vaccination, intensive care	Health, hospital, health system, hospitals, vaccination, national system, system
4	Freedom, democratic, tolerance, against, racism, order, stereotypes, extremism, enemy, ethos	Prime Minister, Antonis, democracy, privileged, national, interests, simply, EAB, presidency	*que, de, en, la, el*, Samaras, democracies
5	Clearance, obligated, disastrous, mouth, entrance, writing, stock market, similar, apportion	Minimum wage, primary, new, CAP, minimum, livestock farmers, restaurant industry, businesses	Farmers, agricultural, products, agricultural, primary, sector, products, disability

Evaluation

Topic	ProdLDA	BERTopic+	BERTopic
1	Obligated, clearance, entrance, minute, overcharge, mouth, enumerate, analyst, objectively	Government, solution, state, reforms, debt, you know, national	*democracy, reforms*, euro, region, development, European, agreement
2	Energy, presidency, directive, regulation, ultimate, ready, multi-year, adoption, migration, unemployment	Stability, memorandums, government, parliament, reforms, debt, Greek people	Tsipras, elections, Mitsotakis, SYRIZA, euro, Mr. Mitsotakis, ladies gentlemen
3	Partner, negotiation, regulation, radical, administration, suffocating, institutional, eurozone, claiming, tax	Health system, doctors, nurses, beds, ICU, hospitals, health, Vasilis, vaccination, intensive care	Health, hospital, health system, hospitals, vaccination, national system, system
4	Freedom, democratic, tolerance, against, racism, order, stereotypes, extremism, enemy, ethos	Prime Minister, Antonis, democracy, privileged, national, interests, simply, EAB, presidency	*que, de, en, la, el*, Samaras, democracies
5	Clearance, obligated, disastrous, mouth, entrance, writing, stock market, similar, apportion	Minimum wage, primary, new, CAP, minimum, livestock farmers, restaurant industry, businesses	Farmers, agricultural, products, agricultural, primary, sector, products, disability

Evaluation

Topic	ProdLDA	BERTopic+	BERTopic
1	Obligated, clearance, entrance, minute, overcharge, mouth, enumerate, analyst, objectively	Government, solution, state, reforms, debt, you know, national	*democracy, reforms*, euro, region, development, European, agreement
2	Energy, presidency, directive, regulation, ultimate, ready, multi-year, adoption, migration, unemployment	Stability, memorandums, government, parliament, reforms, debt, Greek people	Tsipras, elections, Mitsotakis, SYRIZA, euro, Mr. Mitsotakis, ladies gentlemen
3	Partner, negotiation, regulation, radical, administration, suffocating, institutional, eurozone, claiming, tax	Health system, doctors, nurses, beds, ICU, hospitals, health, Vasilis, vaccination, intensive care	Health, hospital, health system, hospitals, vaccination, national system, system
4	Freedom, democratic, tolerance, against, racism, order, stereotypes, extremism, enemy, ethos	Prime Minister, Antonis, democracy, privileged, national, interests, simply, EAB, presidency	*que, de, en, la, el*, Samaras, democracies
5	Clearance, obligated, disastrous, mouth, entrance, writing, stock market, similar, apportion	Minimum wage, primary, new, CAP, minimum, livestock farmers, restaurant industry, businesses	Farmers, agricultural, products, agricultural, primary, sector, products, disability

Evaluation

Topic	ProdLDA	BERTopic+	BERTopic
1	Obligated, clearance, entrance, minute, overcharge, mouth, enumerate, analyst, objectively	Government, solution, state, reforms, debt, you know, national	*democracy, reforms*, euro, region, development, European, agreement
2	Energy, presidency, directive, regulation, ultimate, ready, multi-year, adoption, migration, unemployment	Stability, memorandums, government, parliament, reforms, debt, Greek people	Tsipras, elections, Mitsotakis, SYRIZA, euro, Mr. Mitsotakis, ladies gentlemen
3	Partner, negotiation, regulation, radical, administration, suffocating, institutional, eurozone, claiming, tax	Health system, doctors, nurses, beds, ICU, hospitals, health, Vasilis, vaccination, intensive care	Health, hospital, health system, hospitals, vaccination, national system, system
4	Freedom, democratic, tolerance, against, racism, order, stereotypes, extremism, enemy, ethos	Prime Minister, Antonis, democracy, privileged, national, interests, simply, EAB, presidency	*que, de, en, la, el*, Samaras, democracies
5	Clearance, obligated, disastrous, mouth, entrance, writing, stock market, similar, apportion	Minimum wage, primary, new, CAP, minimum, livestock farmers, restaurant industry, businesses	Farmers, agricultural, products, agricultural, primary, sector, products, disability

Conclusions & Q/A

- Classic algorithms can compare to SOTA
  - Preprocessing is key
  - Especially in multilingual tasks, tailored preprocessing is necessary
- Coherence is incoherent; quantitative evaluation is unreliable in Topic Modeling
  - We optimized based on CV Coherence; in retrospect it's unreliable
  - Some metrics can be used as general filters for really bad models*
- Even SOTA methods require extra steps, such as representation fine-tuning, so as to produce understandable topics

Thank you for your Attention[1]

Nikolaos Giannakopoulos
Dionysios Rigatos

Detailed report of our work and experiments is publicly available on
github.com/DionGR/greek-pm-topic-modeling

[1] (Bahdanau et al., 2016)