GRAD-E1394-F-2023

DECEMBER 2023

Topic Modelling with BERTopic

CARMEN GARRO | FERNANDA ORTEGA | GRESA SMOLICA

KEY DISCUSSION POINTS

Topic modelling with transformers

- More on BERTopic
- Analysis
- Dataset
- Findings
- Additions

Limitations

Sources

TOPIC MODELLING WITH TRANSFORMERS

POLICY RELEVANCE

- Generating information from a different perspective on the current picture of situations of interest.
- Topic modeling: time saving and low cost technique for identifying the underlying pattents in a considerable amount of documents
- Identifying emerging trends --> effective policymaking
- Understanding public sentiment --> tailor made policy measures
- Early detection of potential crisis --> rapid and effective response
- Transparency in intuition and steps of the algorithm

THIS PARTICULAR TUTORIAL

- Focuses on the use of BERTopic to classify a set of news articles that are relevant in the politics and policy areas.
- BERTopic is a technique that uses modularity so that each step can be modified to best fit the problem in question.

MORE ON BERTOPIC - GENERIC PIPELINE

EMBEDDINGS	Documents --> numerical representations
DIM. REDUCTION	Embeddings can be very high-dimensional UMAP-PCA-SVD
CLUSTERING	Similar documents or embeddings are grouped together HDBSCAN - K-means - agglomerative clustering
VECTORIZER	Topic representations --> topic-term matrices Remove stopwords, unfrequent words, etc.
WEIGHTING SCHEME	Considers differences in documents and finds important terms in topics: c-TF-IDF, bm25_weighting
REPRESENTATION TUNNING	Output: representation of the topic KeyBertInspired, PartsOfSpeech-MaximalMarginalRelevance
LLM & GENERATIVE AI	Fine-tuning: generating labels, summaries, poems of topics, etc.

DATASET

- Global News Dataset from Kaggle (Kumar Saksham, 2023)
- Sourced from the NewsAPI - news aggregation service
- More than 100,000 news articles from more than 2,000 media sources, published in the period 01.10.2023 - 29.11.2023
- Selected 9 categories to explore current geopolitical dynamics

PRE-PROCESSING

- Method 1: only removing stop-words
- Method 2: running the model in the raw data
- **Method 3: removing stop-words, lemmatization, stemming, removing numbers, lowercasing, removing punctuation, etc. and then running the model**

REASON WHY

- The nature of the documents, the model picking up unusual words related to the topics, Topic -1 picking up the most generic words but without preprocessing these words spread througout topics

ANALYSIS

Embedding Documents

```
embedding_model = SentenceTransformer('all-MiniLM-L6-v2')
```

UMAP

```
umap_model = UMAP(
    n_neighbors=15,
    n_components=5,
    min_dist=0.05,
    random_state=100)
```

Clusters

```
hdbscan_model = HDBSCAN(
    min_cluster_size=80,
    min_samples=40,
    gen_min_span_tree=True,
    prediction_data=True)
```

ANALYSIS

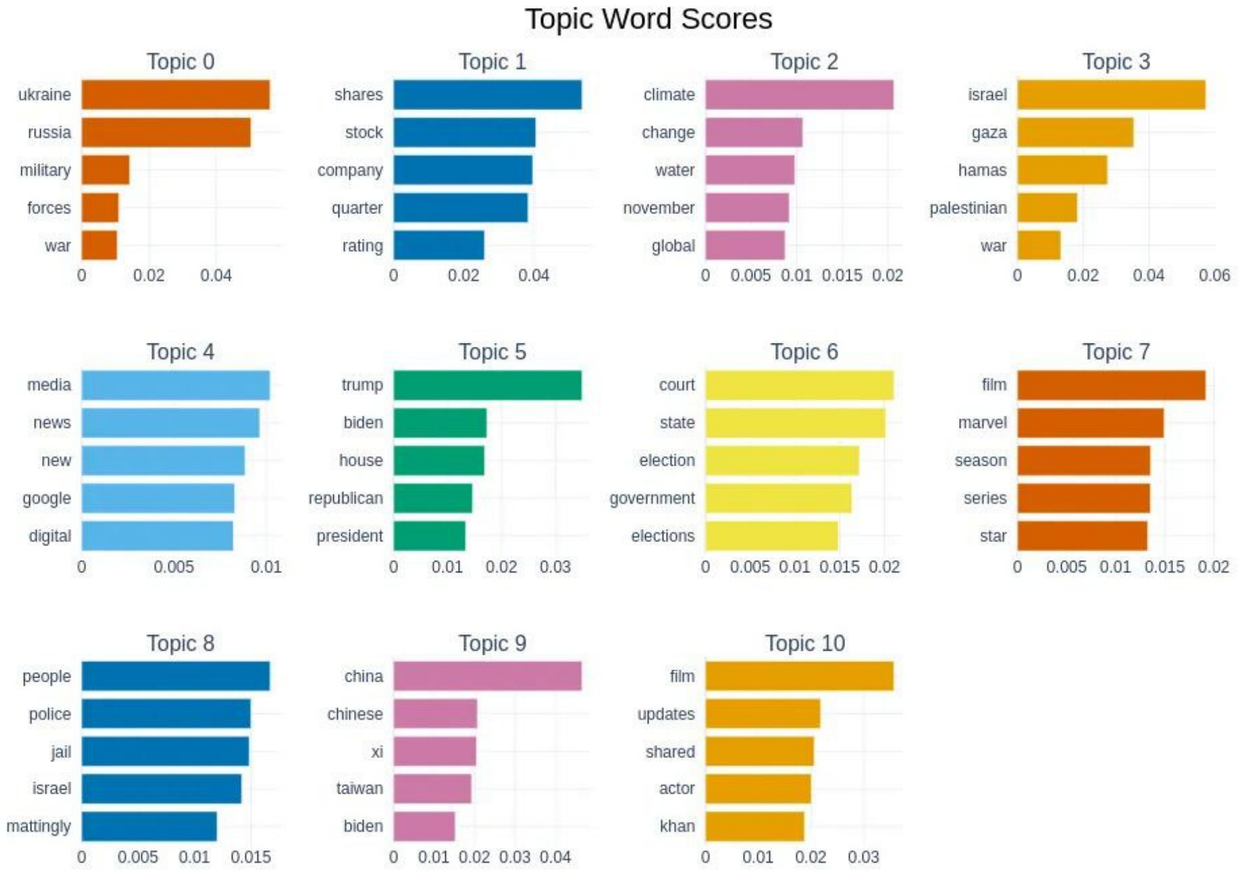
Count Vectorizer

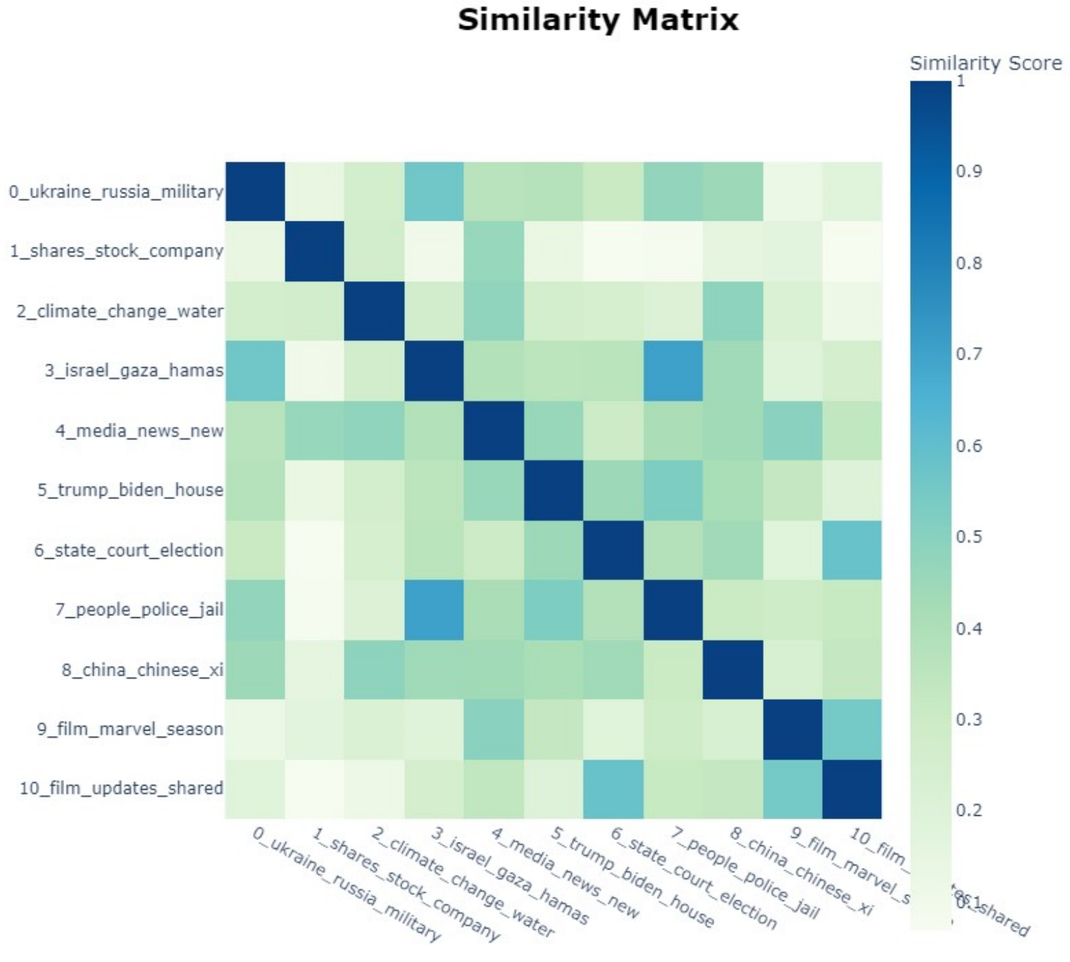
```
vectorizer_model = CountVectorizer(ngram_range=(1, 2),
                                   stop_words=stopwords)
```

Model

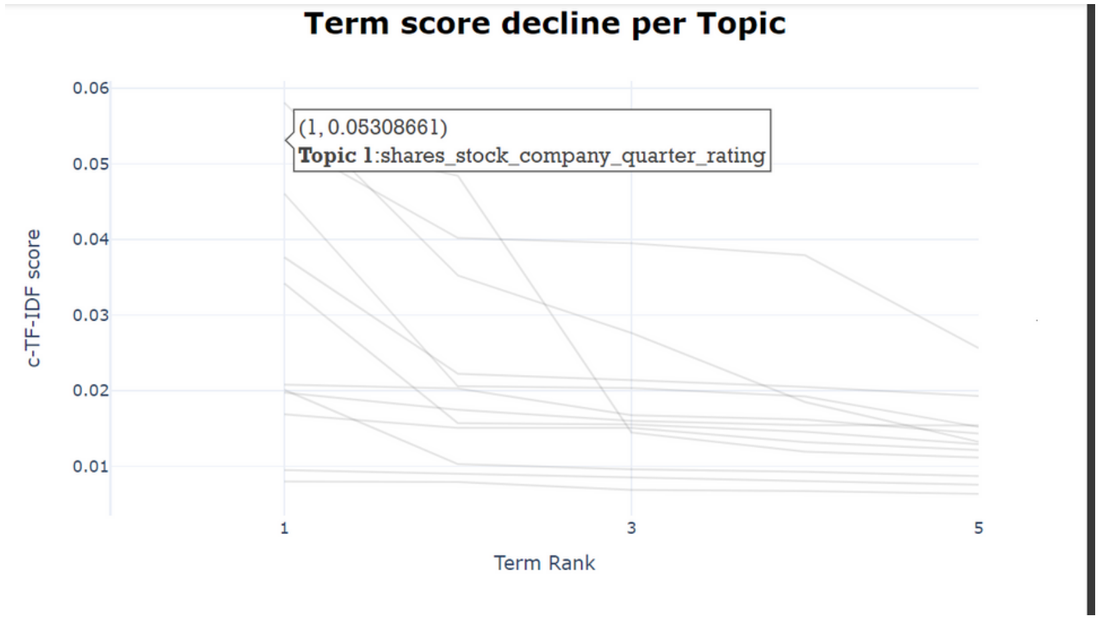
```
model = BERTopic(
    umap_model=umap_model,
    hdbscan_model=hdbscan_model,
    embedding_model=embedding_model,
    vectorizer_model=vectorizer_model,
    top_n_words=5,
    language='english',
    calculate_probabilities=True,
    verbose=True
)
topics, probs = model.fit_transform(df['full_content'])
```

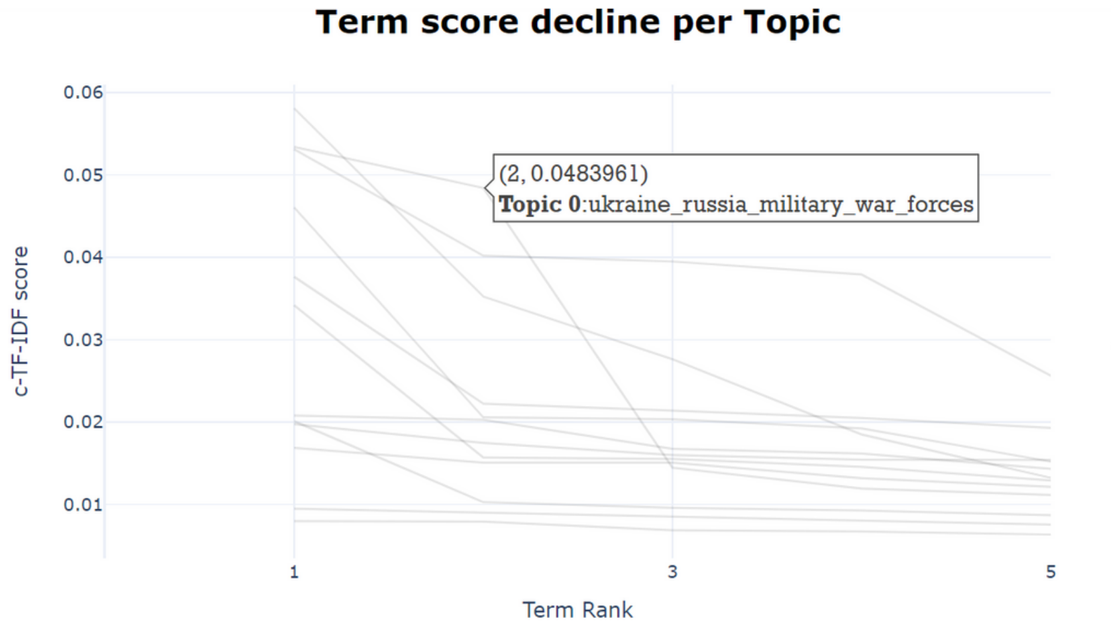
FINDINGS


Topic Word Scores
Topic 0
ukraine
russia
military
forces
war
0
0.02
0.04
Topic 1
shares
stock
company
quarter
rating
0
0.02
0.04
Topic 2
climate
change
water
november
global
0
0.005
0.01
0.015
0.02
Topic 3
israel
gaza
hamas
palestinian
war
0
0.02
0.04
0.06
Topic 4
media
news
new
google
digital
0
0.005
0.01
Topic 5
trump
biden
house
republican
president
0
0.01
0.02
0.03
Topic 6
court
state
election
government
elections
0
0.005
0.01
0.015
0.02
Topic 7
film
marvel
season
series
star
0
0.005
0.01
0.015
0.02
Topic 8
people
police
jail
israel
mattingly
0
0.005
0.01
0.015
Topic 9
china
chinese
xi
taiwan
biden
0
0.01
0.02
0.03
0.04
Topic 10
film
updates
shared
actor
khan
0
0.01
0.02
0.03

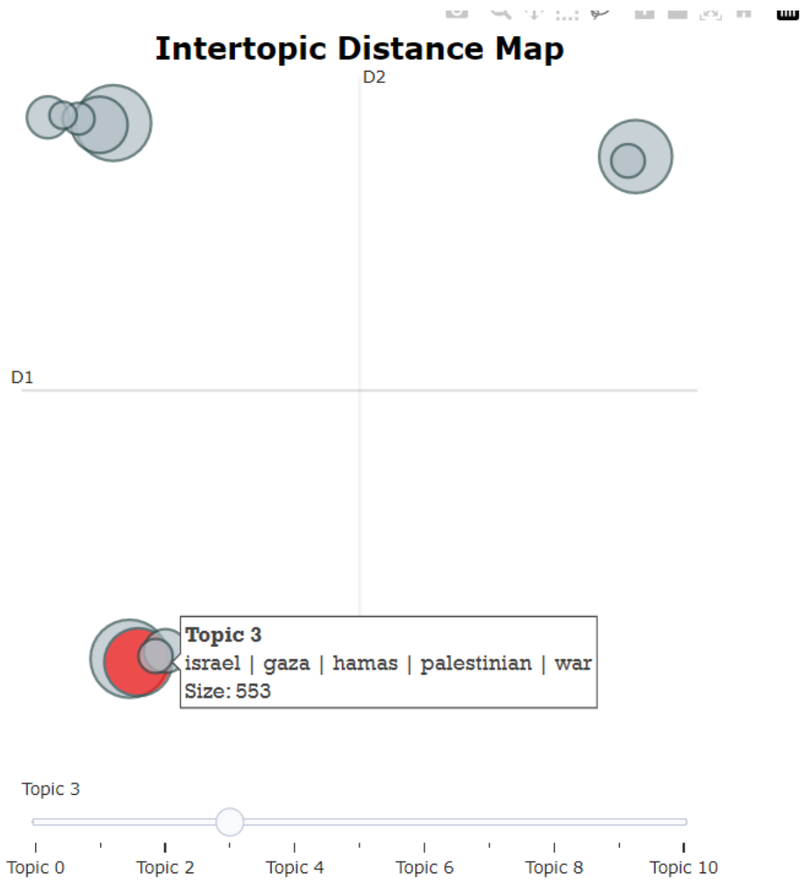

Similarity Matrix
Similarity Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0_ukraine_russia_military
1_shares_stock_company
2_climate_change_water
3_israel_gaza_hamas
4_media_news_new
5_trump_biden_house
6_state_court_election
7_people_police_jail
8_china_chinese_xi
9_film_marvel_season
10_film_updates_shared

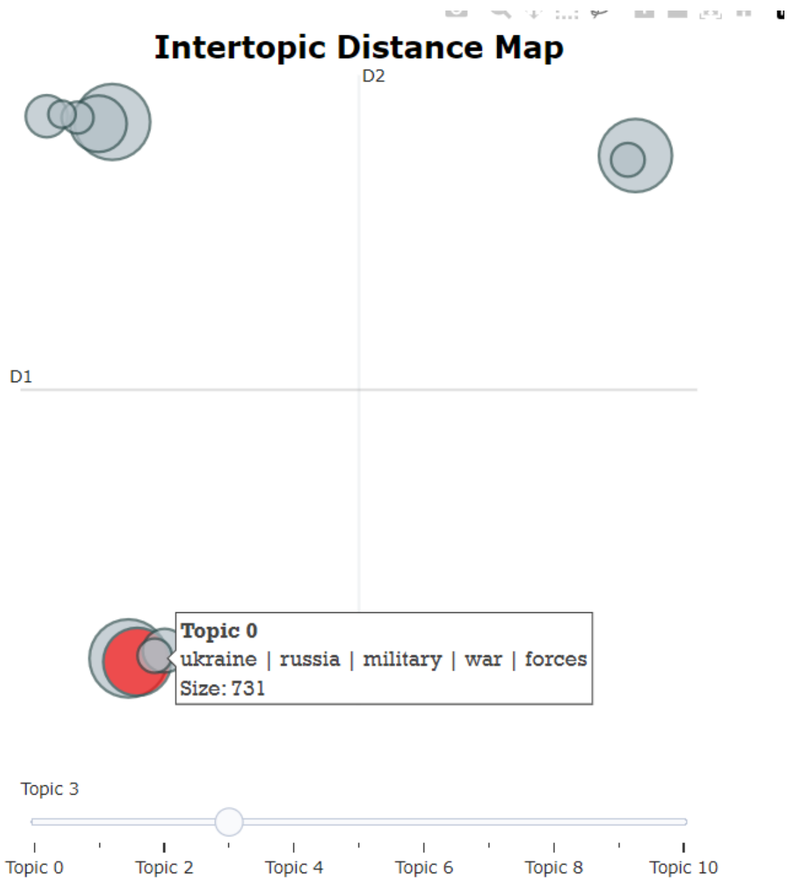
FINDINGS


Term score decline per Topic
0.06
0.05
0.04
0.03
0.02
0.01
c-TF-IDF score
(1, 0.05308661)
Topic 1:shares_stock_company_quarter_rating
1
3
5
Term Rank


Term score decline per Topic
0.06
0.05
0.04
0.03
0.02
0.01
c-TF-IDF score
(2, 0.0483961)
Topic 0:ukraine_russia_military_war_forces
1
3
5
Term Rank

FINDINGS


Intertopic Distance Map
D2
D1
Topic 3
israel | gaza | hamas | palestinian | war
Size: 553
Topic 3
Topic 0
Topic 2
Topic 4
Topic 6
Topic 8
Topic 10


Intertopic Distance Map
D2
D1
Topic 0
ukraine | russia | military | war | forces
Size: 731
Topic 3
Topic 0
Topic 2
Topic 4
Topic 6
Topic 8
Topic 10

WHAT'S NEXT UNTIL NEXT WEEK?

COHERENCE METRICS

- Degree of significance between the words inside a topic in terms of how interpretable it is
  - C_v: higher score --> better understandable topic by a human
  - C-umass: the closer to zero --> the better

SENTIMENT ANALYSIS

- Sentiment within selected categories

EXPLORING THE GENERATIVE COMPONENT OF BERTOPIC

- Experimenting with article generation (if feasible)

LIMITATIONS

- Further exploration of theoretical approaches and possibilities would be useful because of the amount of possible tweaks in this model
- Uneven number of articles within categories this can have several effects:
  - bias towards overrepresented categories
  - difficulty in identifying less common topics
  - overfitting of dominant categories
  - resampling could be used to level the data
- As a general limitation, BERTopic works better with larger datasets
- Documents are assumed to only have one topic
- Computational resources: this model takes around 20 mminutes to run

SOURCES

- Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., & Buntine, W. (2021). Topic modelling meets deep neural networks: A survey. arXiv preprint arXiv:2103.00498.
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.
- Jin, M., Luo, X., Zhu, H., & Zhuo, H. H. (2018). Combining deep learning and topic modeling for review understanding in context-aware recommendation. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Presented at the Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana. doi:10.18653/v1/n18-1145
- Karlsson, M., & Sjøvaag, H. (Eds.). (2018). Rethinking Research Methods in an Age of Digital Journalism (1st ed.). Routledge. https://doi.org/10.4324/9781315115047

SOURCES

- Konstantina Andronikou. (2022, October 21). Topic Modeling with BERTopic—The Analytics Lab. Https://Www.Theanalyticslab.Nl/. https://www.theanalyticslab.nl/topic-modeling-with-bertopic/
- Briggs, J. (n.d.). Advanced Topic Modeling with BERTopic | Pinecone. Retrieved December 6, 2023, from https://www.pinecone.io/learn/bertopic/xx
- Kumar Saksham. (n.d.). Global News Dataset. Retrieved December 6, 2023, from https://www.kaggle.com/datasets/everydaycodings/global-news-dataset
- Topic Modeling with Deep Learning Using Python BERTopic. (2023, March 28). Medium. https://medium.com/grabngoinfo/topic-modeling-with-deep-learning-using-python-bertopic-cf91f5676504
- Kumar Saksham. (2023). Global News Dataset [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/7105651