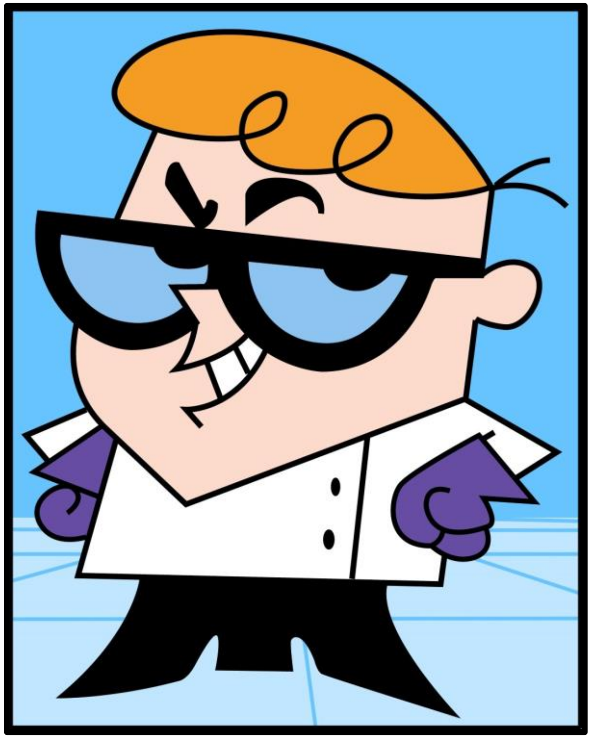
OTMISC: Our Topic Modeling Is Super Cool

Technische Universität München

Fakultät für Informatik

NLP Lab Course, SS22

26.07.2022



Berk Sudan, Ferdinand Kapl, Yuyin Lang

Topic Modeling Advancements

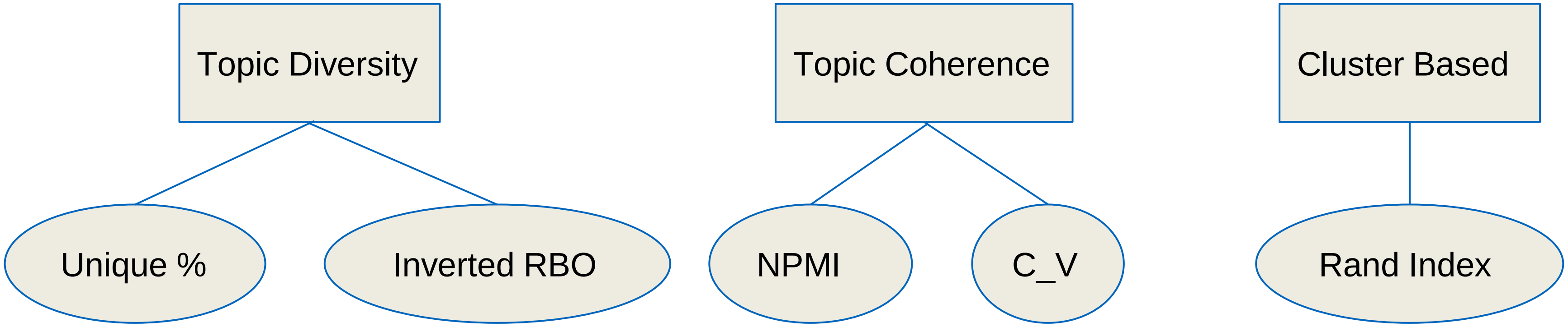
Motivation

Motivation

- **Goal:** Identify topics in large unstructured text data (documents)
- **Method:** Cluster documents and associate topic words
  - Old Approach: Probability based
  - Advancements: Embedding based
- **Use Cases:**
  - Recommender systems
  - Recruiting algorithms
  - Organize Emails / Customer reviews / Social Media profiles

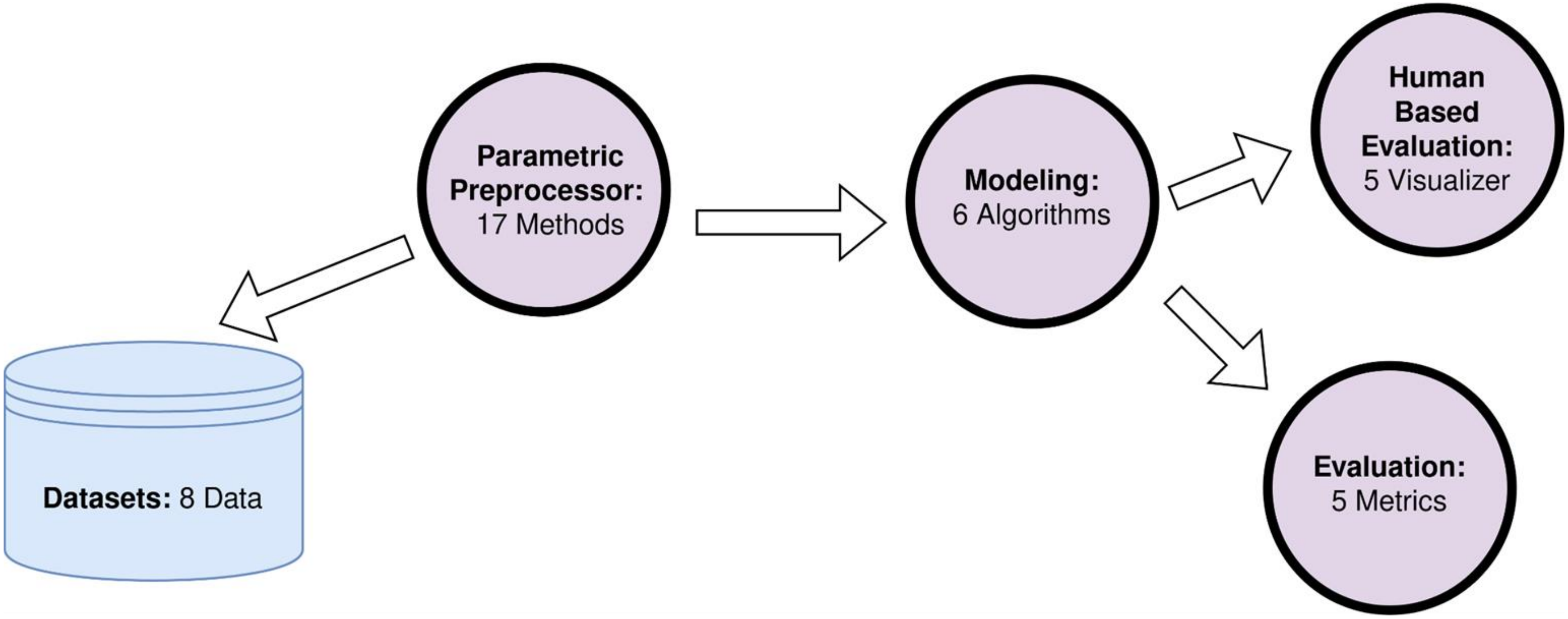
Related Work

- Topic Modeling Advancements: Top2Vec, BERTopic, CTM, LDA-BERT…
- **Issue:** No golden standard for evaluation of topic models
- **Solution:** Exhaustive combination of popular used metrics & human based evaluation based on visualizations



Architectural Design

OTMISC Architectural Design



Datasets

Available Datasets

- Currently: **8 Datasets**
- **Long Text Datasets**
  - 20 News
  - Yahoo Answers
  - AG News - News Text
- **Short Text Datasets**
  - AG News - News Title
  - Crisis Resource 1,7,12,17

Short Text Data Example (Crisis 12)

Tweets
RT @diplo: Twerkbook pro #plurnt #earthquake http://t.co/5x5ya6wxF6
In @BBCUrdu #Balochistan EarthQuake - program NDMA's Military Official Continue to Refuse accepting outside help http://t.co/Xs8wK4glP7 …
RT @ErumManzoor: People who wanna help #earthquake affectees in Baluchistan can contact @AsimBajwaISPR

- Tweets
- Often less than 30 words
- Contains: URLs, Hashtags, Typos

Long Text Data Example (20 News)

From games Subject companies in vehicle market Article-I.D Distribution world Lines 34 NNTP-Posting-Host What would all of you out there in net land think of the big General getting together and to study exactly what the market price are for building and say to do that that most of the military for are out of somewhere say has the ever really used that You get the idea figure out how many how often where to etc ... Then taking this data and type company bad example know ... but at least its an example ... To develop between and Then to take all of those and figure out what the are and those in order to that ca n't be built today And say that this again by the cost about 20 million And from here all of these companies went their separate ways with the of taking all of the market data and the design data to and saying ``

- News
- Often 100+ words
- More readable
- No Sparsity

Preprocessing

Preprocessing

- LDA, NMF: Use BOW
- LDA-BERT: LDA part needs BOW

Preprocessing

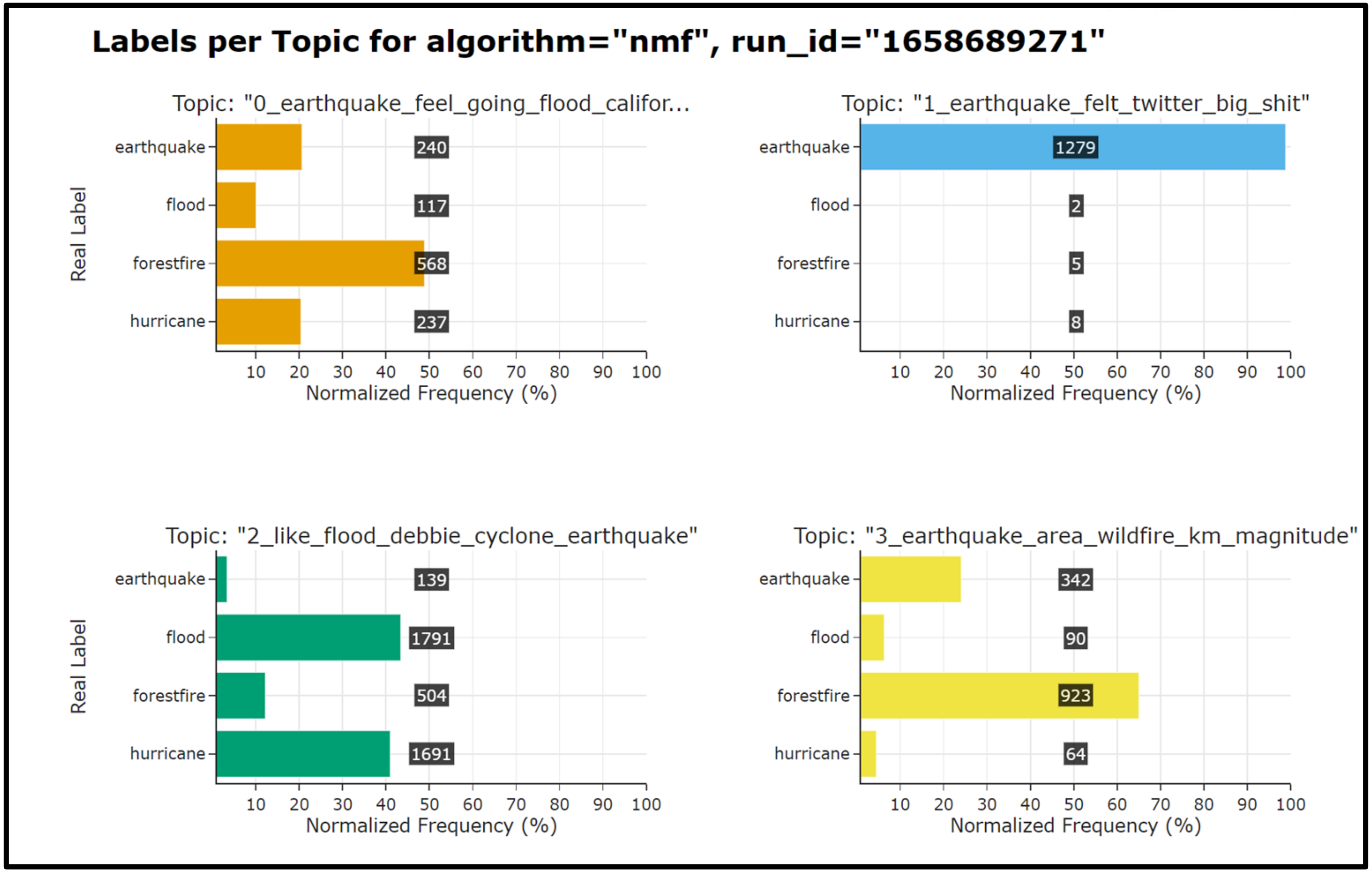
- *The essential preprocessing methods:*
  - Lower case
  - Remove stop words
  - Lemmatize (to noun)

- *Special for Tweets:*
  - Remove url
  - Remove tags

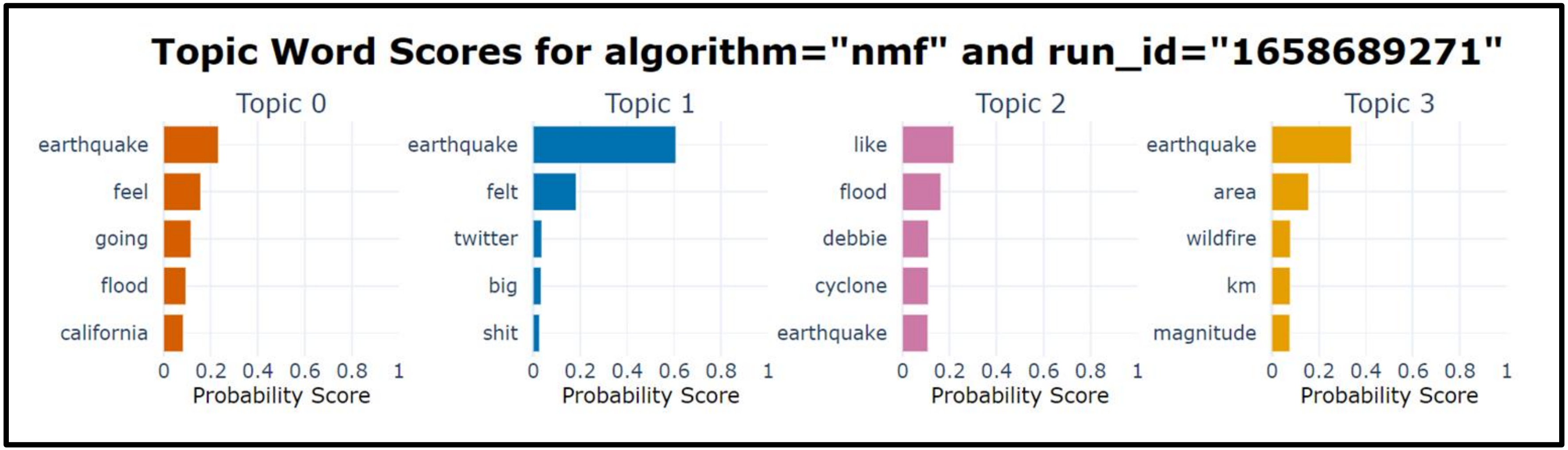
LDA & NMF

Algorithms – LDA & NMF

Crisis_12



Algorithms – LDA & NMF

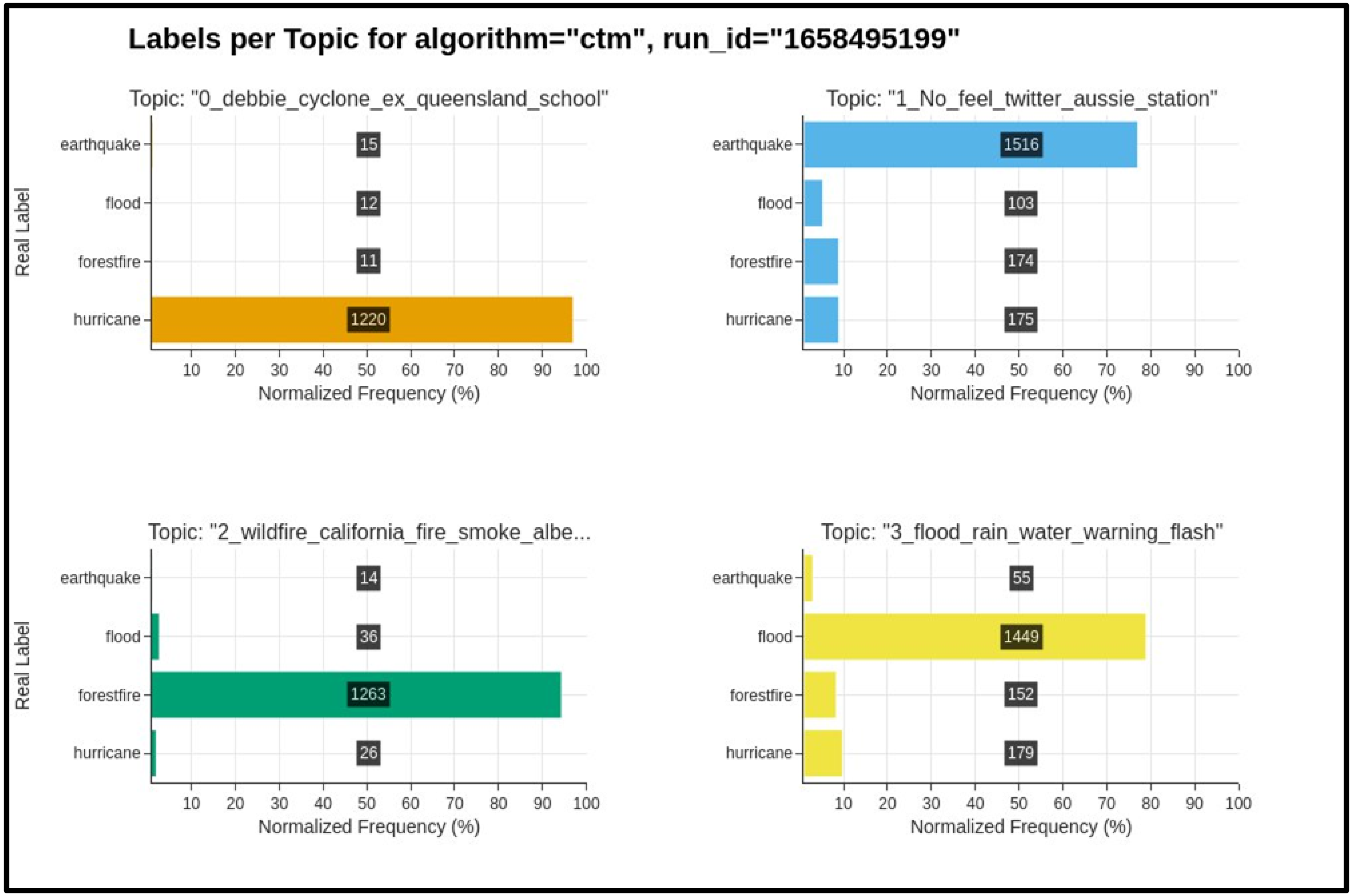


NMF on Preprocessed and Unpreprocessed Data (Diversity Inv. RBO)

Dataset	Preprocessed	Not Preprocessed
20news	0.960	0.822
ag_news_short	**0.904**	**0.698**
crisis_01	0.878	0.740
crisis_07	**0.863**	**0.664**
crisis_12	0.729	0.814
crisis_17	0.936	0.872
yahoo	0.939	0.792

CTM

Algorithms – CTM



Algorithms – CTM

	Document ID	Document	Real Label	Assigned Topic Num	Assignment Score
2499	2499	expect flood state law change follow maybe business clear	flood	2	0.766885
2580	2580	flash flood warning right careful driving people	flood	2	0.725889
1837	1837	found truck flood plain hell rain know drowned	flood	2	0.689359
2861	2861	nepal flood wake call response	flood	2	0.679502
2828	2828	flood event wind event number car houston area	flood	2	0.678766

Top2Vec

Top2Vec - Available Parameters

- **Minimum Topic Words:** Depends on corpus size and its vocabulary.
- **Embedding Model:** Tested 8 Models
- **Umap & Hdbscan Args**
- **Number of Topics:** Hierarchical Topic Reduction

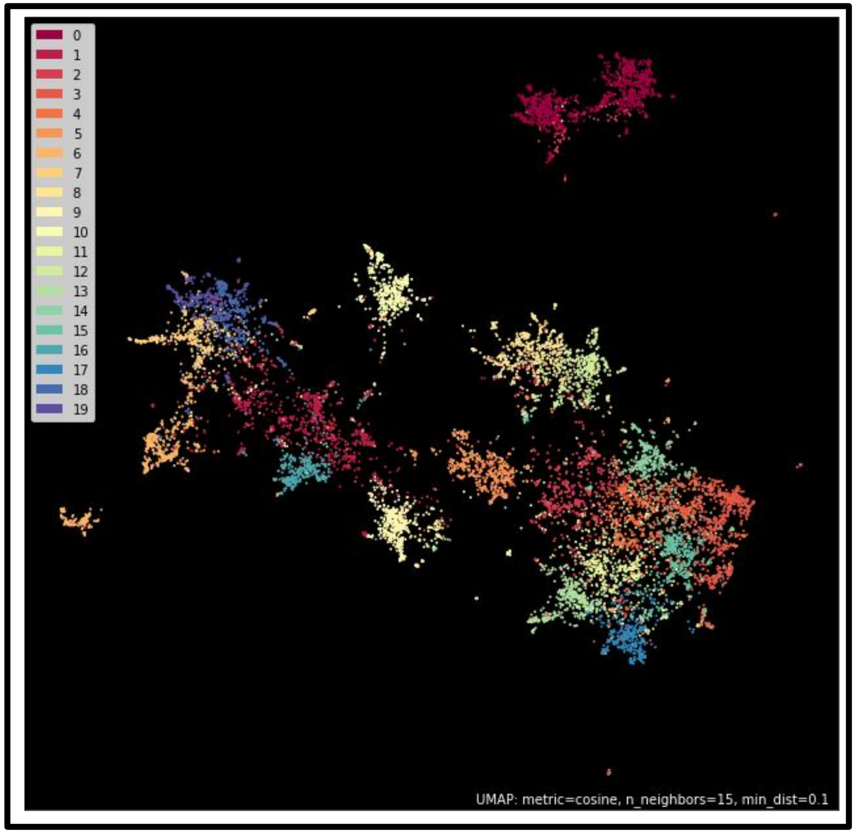
Top2Vec - Topic Assignments with Scores

Document	Real Label	Assigned Topic Num	Assignment Score
Wow just had a earthquake	earthquakes	1	0.914956
Ohhh shit earthquake	earthquakes	1	0.914789
ummm earthquake anyone??	earthquakes	1	0.913467
Holy shit earthquake	earthquakes	1	0.913255
Um earthquake anyone?	earthquakes	1	0.912174

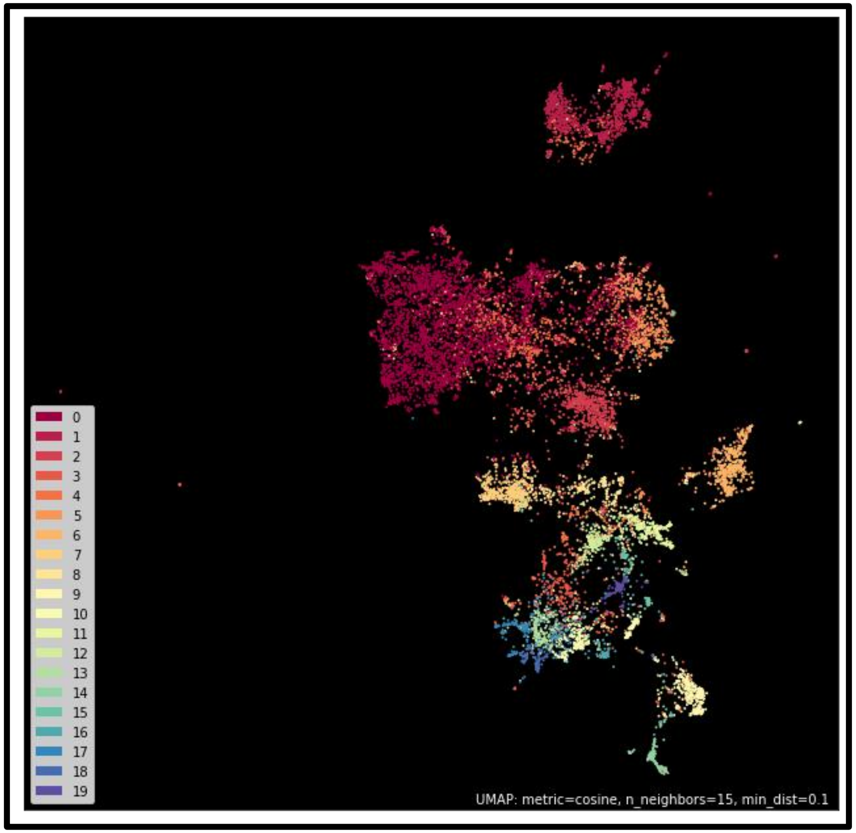
Parameters: Embedding=”universal-sentence-encoder-large”, Data=”CRISIS-12”

Assignment Score: The cosine similarity of the document and topic vector.

Visualization - UMap



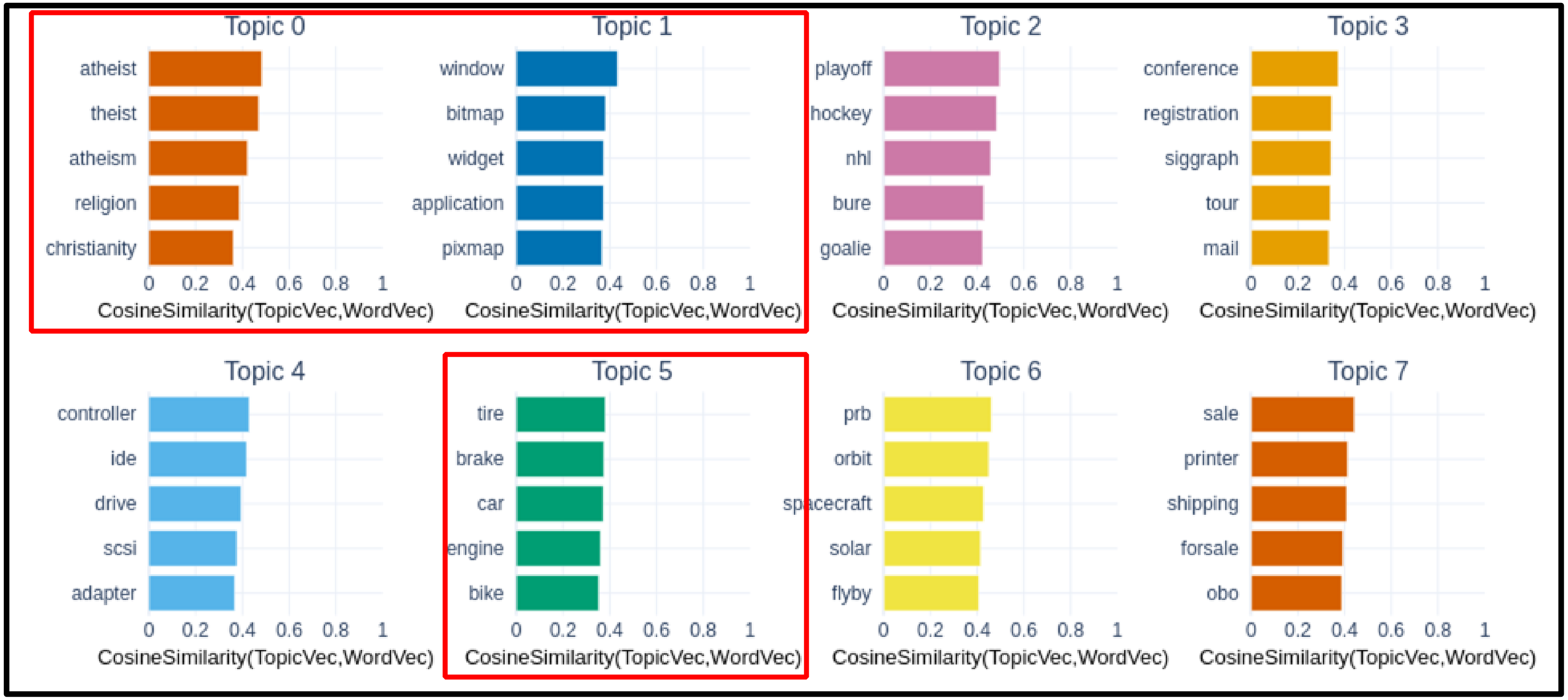
universal-sentence-encoder



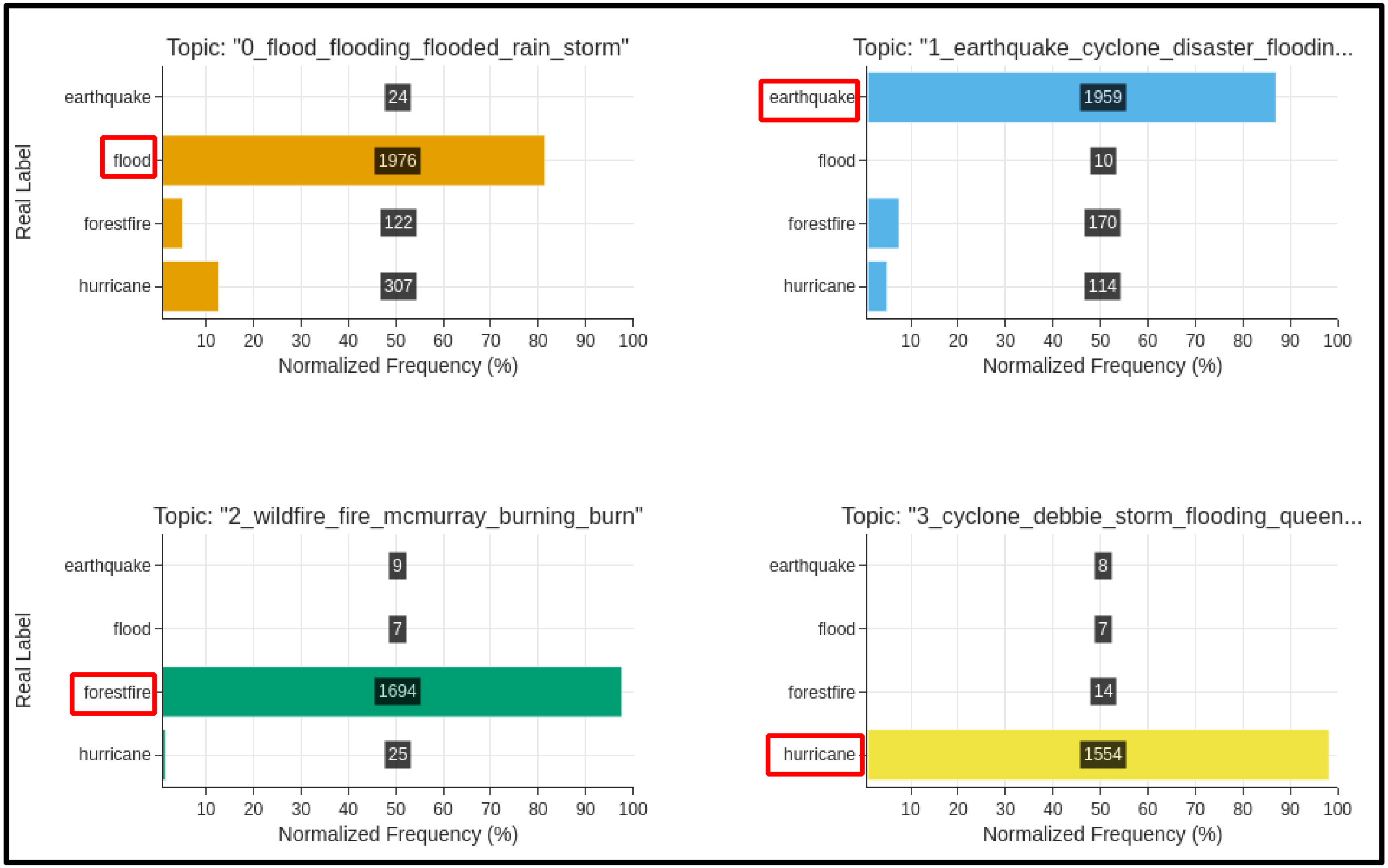
all-MiniLM-L6-v2

Advantage: See outliers

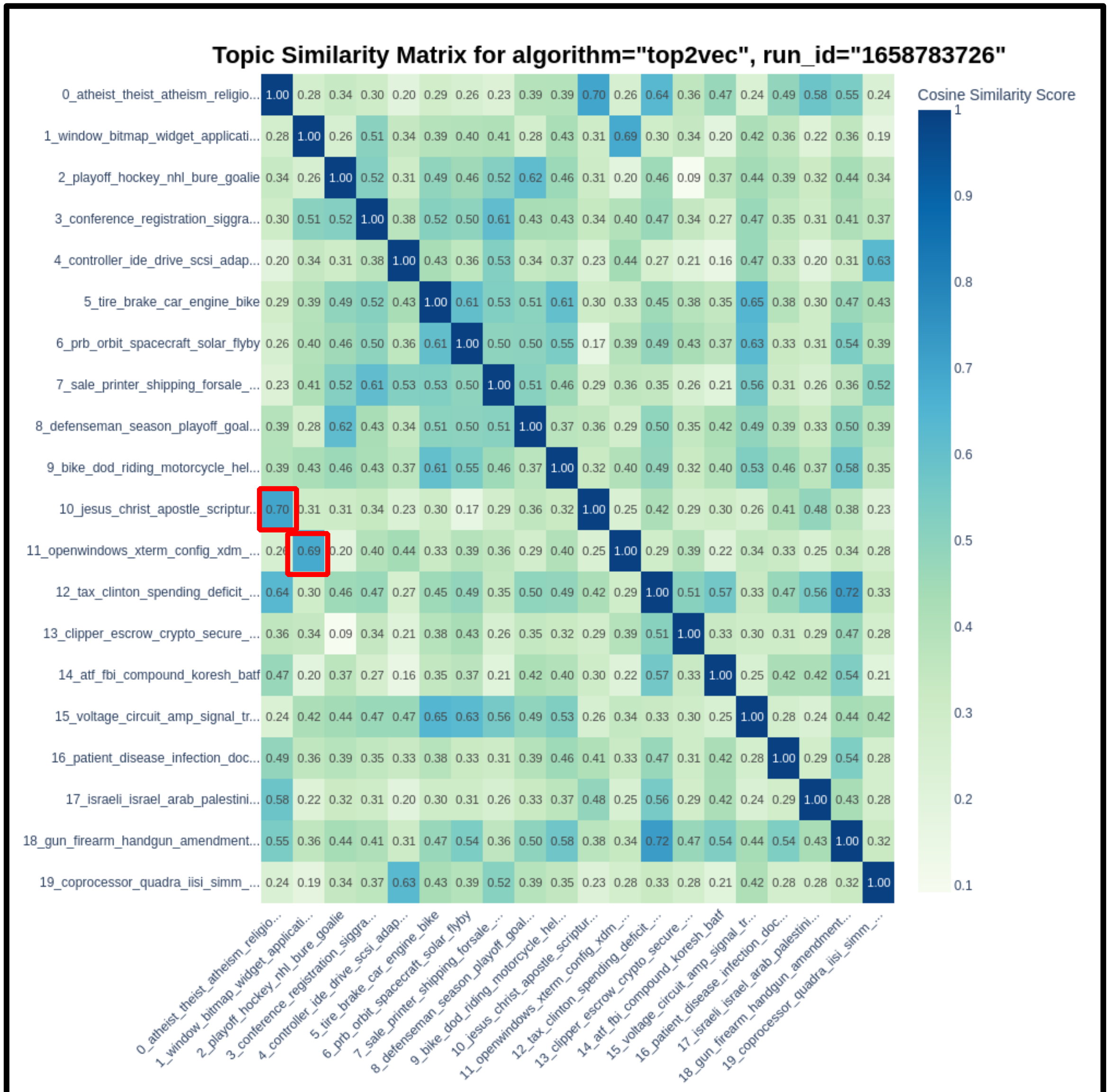
Visualization - Similar Top Topic Words



Visualization - Labels per Topic



Visualization - Topic Similarity Matrix

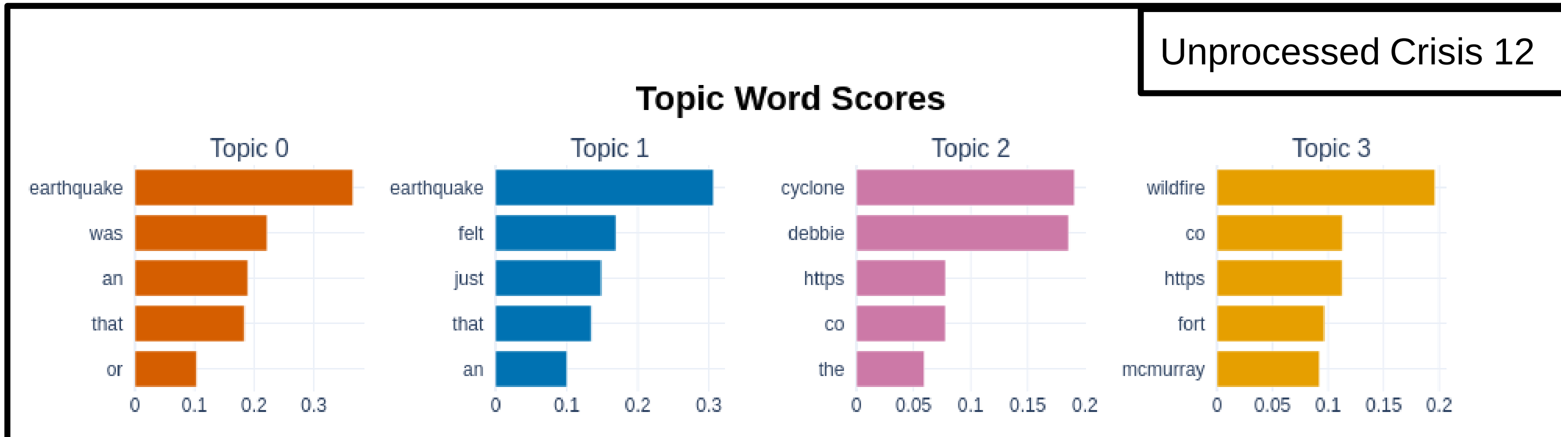


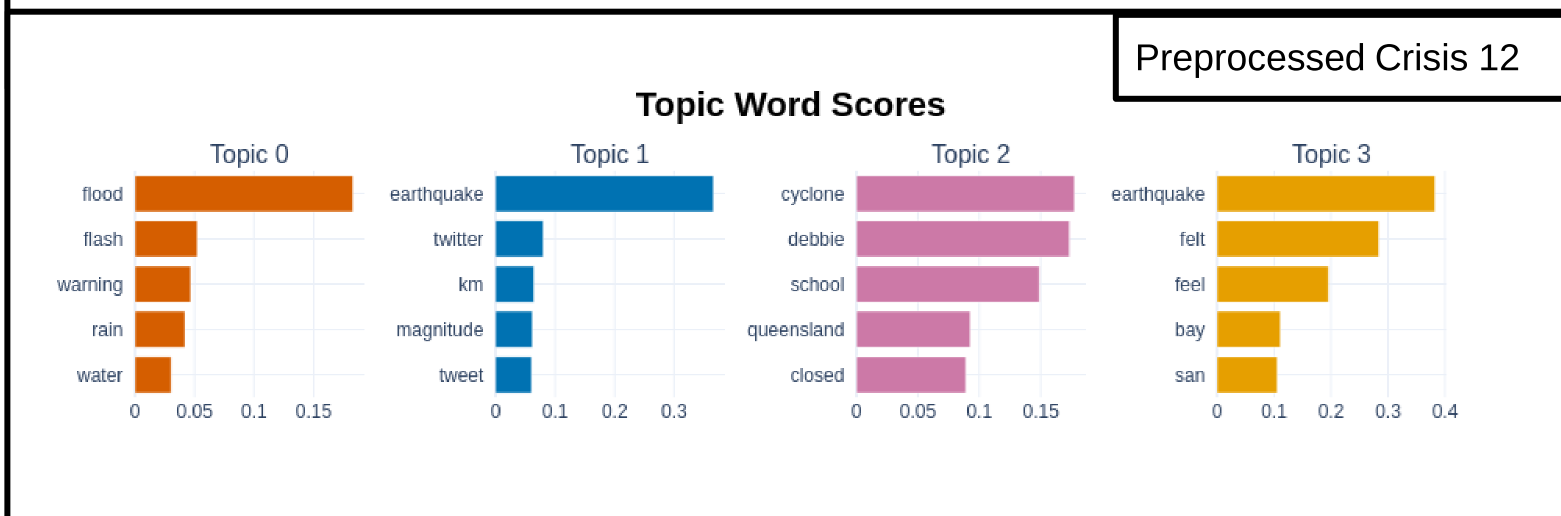
Top2Vec on Different Embedding Models

	TD	TD	TC	TC	Cluster
<u>**Embedding Model**</u>	**Unique**	**Inv. RBO**	**NPMI**	**C_V**	**Rand**
all-MiniLM-L6-v2	0.907	0.957	-0.284	0.388	0.864
doc2vec	0.912	0.931	-0.276	0.485	0.552
paraphrase-multilingual-MiniLM-L12-v2	0.930	0.987	-0.252	0.376	0.777
universal-sentence-encoder	0.860	0.941	-0.267	0.366	0.827

BERTopic & LDA-BERT

BERTopic - Compare Best Results on Crisis 12

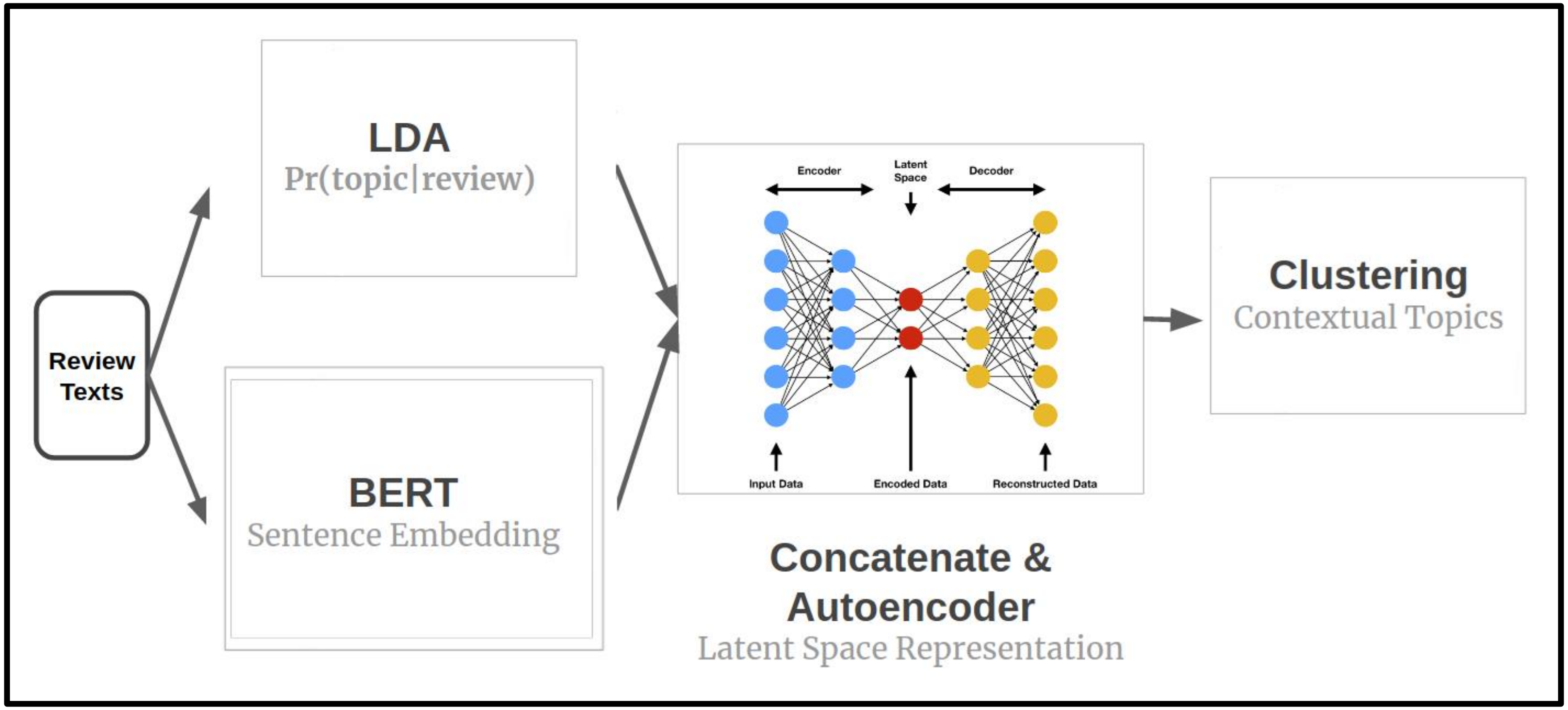




Impression: Topic words for the preprocessed version are more descriptive

LDA-BERT

- **An Upgrade:** LDA + BERT



Source: https://blog.insightdatascience.com/contextual-topic-identification-4291d256a032

Evaluation & Comparisons

Evaluation: Long Text

	TD	TD	TC	TC	Cluster
Algorithm	**Unique**	**Inv. RBO**	**NPMI**	**C_V**	**Rand**
NMF	0.565	0.805	0.035	0.508	0.846
LDA	0.525	0.878	0.038	0.538	0.792
LDA-BERT	0.397	0.661	0.055	**0.594**	0.880
BERTopic	0.637	0.828	**0.081**	0.577	0.425
Top2Vec	**0.819**	0.902	-0.113	0.436	**0.906**
CTM	0.552	**0.909**	-0.056	0.404	0.861

Evaluation: Short Text

	TD	TD	TC	TC	Cluster
Algorithm	**Unique**	**Inv. RBO**	**NPMI**	**C_V**	**Rand**
NMF	0.718	0.715	-0.014	0.388	0.614
LDA	0.700	0.810	0.010	0.435	0.625
LDA-BERT	0.743	0.778	-0.001	0.421	0.727
BERTopic	0.826	0.901	**0.050**	**0.493**	0.474
Top2Vec	0.897	0.956	-0.279	0.388	0.739
CTM	**0.966**	**0.994**	-0.100	0.486	**0.746**

Take-Aways

Take-Aways

- Embedding based models perform better
- Winners on Short Data:
  - CTM & BERTopic
- Winners on Long Data:
  - CTM & Top2Vec

Challenges and future work

- Limited computing resource
- Whole pipeline creation
- LDA-BERT Realization
- Unsupervised
- Slow CTM

- Make the work to a real topic modelling tool (in Github)

Q & A