CW4 - TOPIC MODELING WITH BERT
Matriculation Number - 2549408b

Abstract:

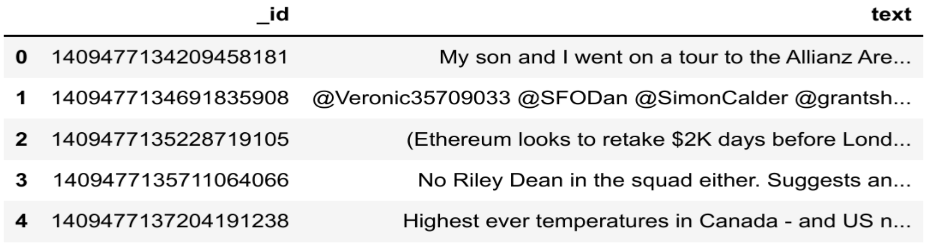
In recent years, as the number of internet users grows rapidly, we are left with large volumes of unstructured text data. Starting from social media posts, emails, blog posts, and so on. To handle this large amount of text data we need a technique to automatically extract the contextual meaning from those documents by spotting the recurrent topics [1]. Topic modeling is the form of unsupervised learning techniques which extract the topics when given a set of documents [2].

Introduction:

Popular topic models like Latent Dirichlet Allocation(LDA) and Non-Negative Matrix Factorization (NMF) could be used as baseline models, while we use transformer-based model BERT since pre-trained models give more accurate representations of words and sentences. Here we are using a library called BERTopic which uses BERT and transformers embeddings to do topic modeling. The BERTopic creates a sentence transformer, let's say we have a document that has a bunch of words or sentences. Each word is converted to word vectors and is then condensed into sentence vectors, so one vector represents one tweet (or doc). This is then going to perform dimensionality reduction followed by the topic clustering. This technique leverages BERT embeddings and a class-based TF-IDF to create dense clusters for easily interpretable topics and spotting keywords for topic descriptions. It gives us an idea of what these documents are collectively talking about. These are the results of the cluster-based tf-idf.

Data:

The Dataset we are using for this topic modeling is twitter data with a total of 53399 tweets with a unique id for each. The goal is to topic model these tweets and clusters them into different categories.



	_id	text
0	1409477134209458181	My son and I went on a tour to the Allianz Are...
1	1409477134691835908	@Veronic35709033 @SFODan @SimonCalder @grantsh...
2	1409477135228719105	(Ethereum looks to retake $2K days before Lond...
3	1409477135711064066	No Riley Dean in the squad either. Suggests an...
4	1409477137204191238	Highest ever temperatures in Canada - and US n...

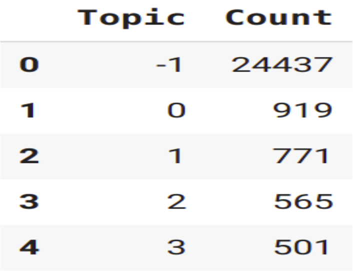
BERTopic Model:

By calling the *BERTopic()* function we can create a model and do *fit_transform* to the text data in our CSV file. There are two outputs generated during the *fit_transform* that are *topics* and *probabilities*. The *topics* represent the topic it is assigned to and *probabilities* show the likelihood of a document falling into any of the possible topics.

Topics can be represented with respect to their relative frequency. In the above frequency, the table generated the topic -1 refers to the outlier topic that does not necessarily belong to any of the topics generated. In this dataset, 24437 documents are not classified indicated it does not belong to any of the clusters. The other 28962 documents are clustered under 595 topics.

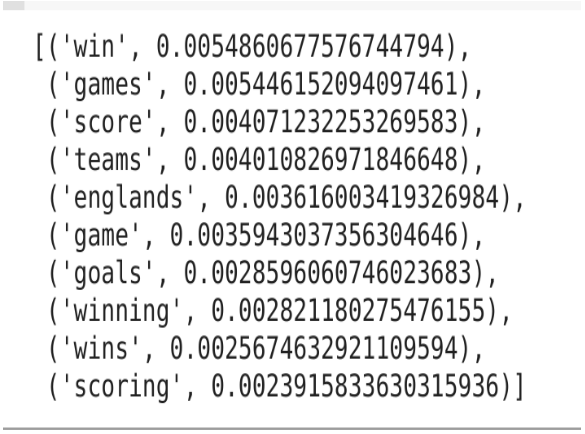
```
24437 documents have not been classified
The other 28962 documents are 595 topics
```

Topic 0 has the most frequent topic with 919 tweets belonging to it.



	Topic	Count
0	-1	24437
1	0	919
2	1	771
3	2	565
4	3	501

If we inspect that topic using *get_topic(),* it gives us the most frequent words in that topic cluster. Thus the interpretation is done for the topic needed. For eg: If we are to inspect topic 0 we get the below words, this gives us the overall context of this cluster and tells us this cluster tells us about sports.

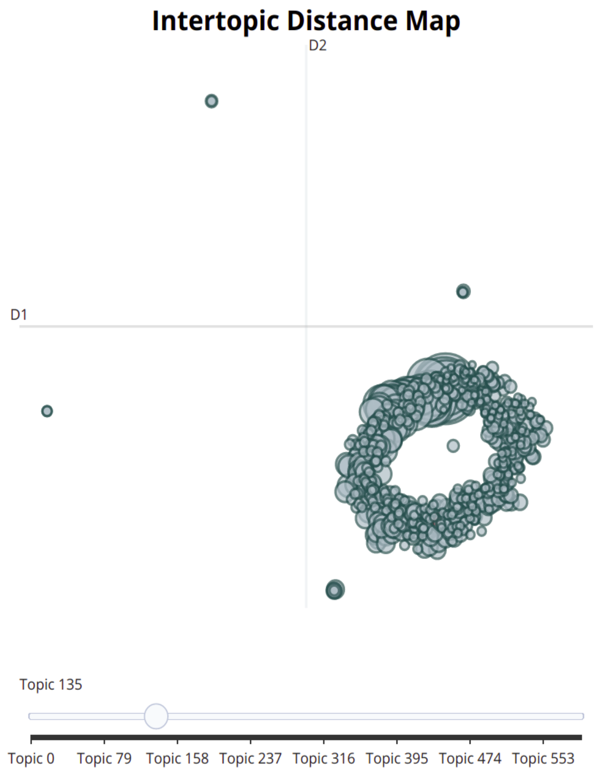


```
[('win', 0.005486067757674479 4),
 ('games', 0.005446152094097461),
 ('score', 0.004071232253269583),
 ('teams', 0.004010826971846648),
 ('englands', 0.003616003419326984),
 ('game', 0.0035943037356304646),
 ('goals', 0.0028596060746023683),
 ('winning', 0.002821180275476155),
 ('wins', 0.0025674632921109594),
 ('scoring', 0.0023915833630315936)]
```

BERTopic supports different *languages* and also supports multilingual documents in the dataset. A pre-trained embedding model could be passed to the BERTopic with the variable *embedding_model.*

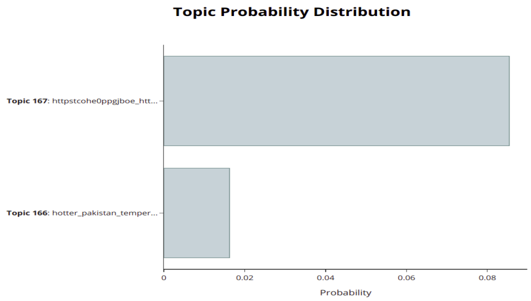
Visualization Topics:

BERTopic gives us similar interactive visualization just like LDAvis, which allows us to explore topics and the words that describe them. BERTopic uses embedded class-based TF-IDF representation of the topics in 2D using Umap. Then, it visualizes the dimensions using Plotly to give an interactive view. In this visualization, each circle indicates a topic and its size is the frequency of the topic across all documents.



Visualize Probabilities:

BERTopic allows us to visualize the probability of that document belonging to each possible topic. We just call *visualize_distribution* to understand the probability distribution of the most probable topics for that particular document.

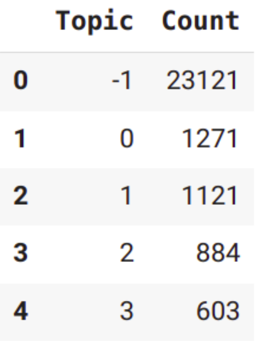


Topic Reduction:

According to the above visualization over 500 topics have been generated. It is a bit fine-grained solution. So the next step, to optimize our BERTopic model is to reduce the number of topics by merging pairs of topics that are most similar to each other, as indicated by the cosine similarity between c-TF-IDF vectors. If we have prior knowledge about the possible number of topics in the dataset, we can explicitly set the variable *nr_topics* and train the model. If that’s not the case, then we can do Automatic Topic Reduction by setting the *nr_topics=” auto”* will force to merge the pair of topics is found that exceeds a minimum similarity of 0.9.

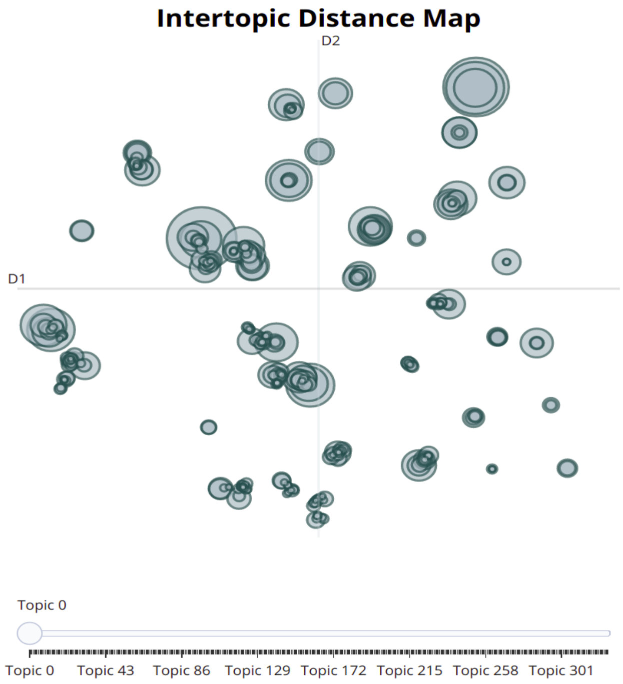
But doing this fine-tuning we are getting the following results

```
23121 documents have not been classified
The other 30278 documents are 328 topics
```

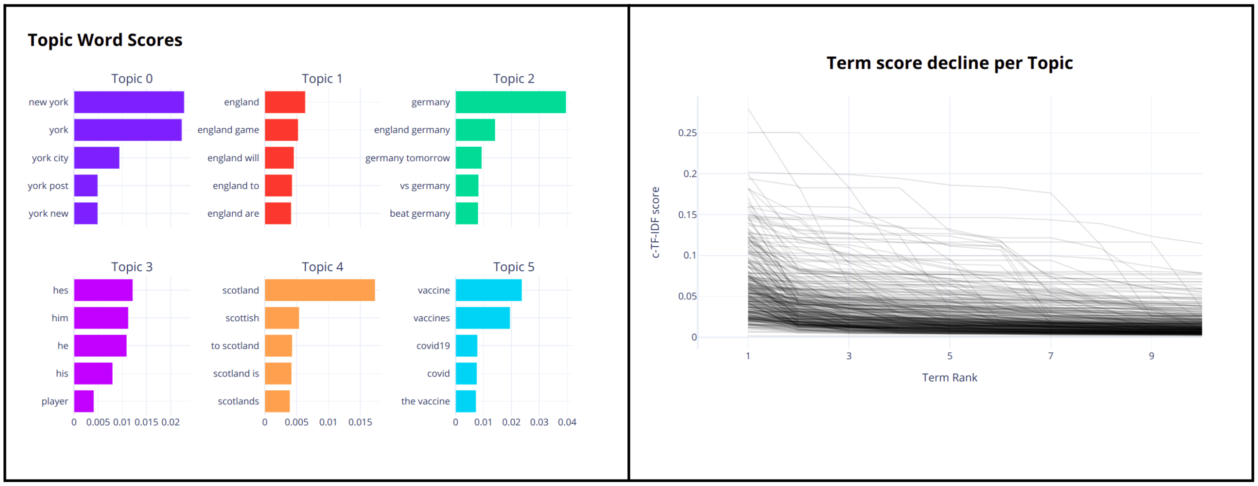


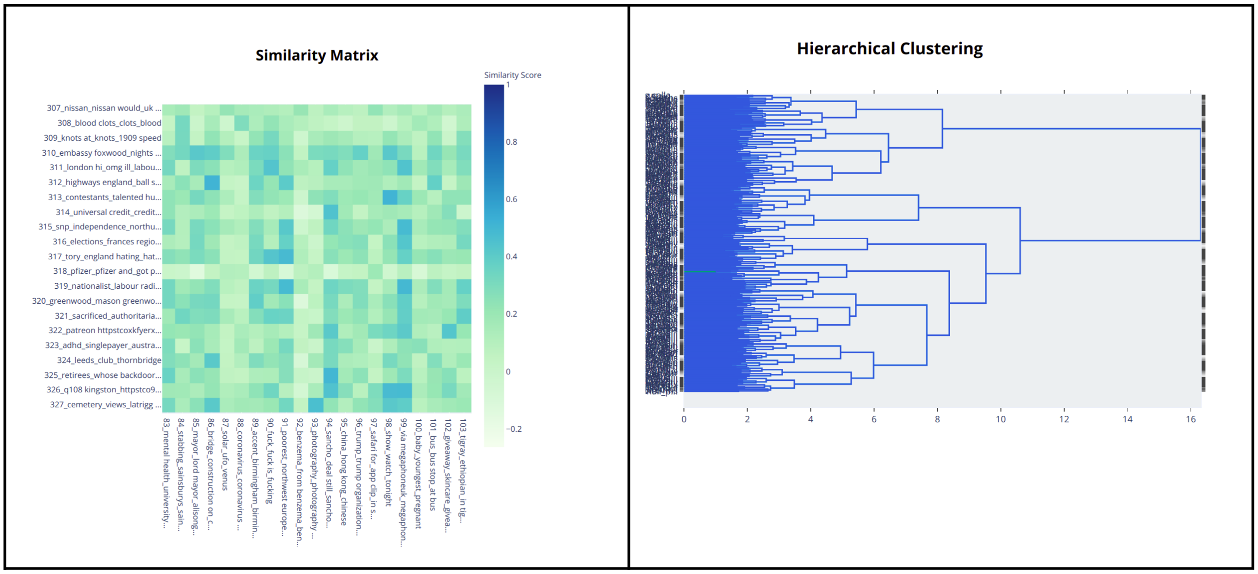
	Topic	Count
0	-1	23121
1	0	1271
2	1	1121
3	2	884
4	3	603

The number of outliers is reduced and around 30278 are been classified into 328 topics. The number of topics has been significantly reduced. The final topic visualization looks like

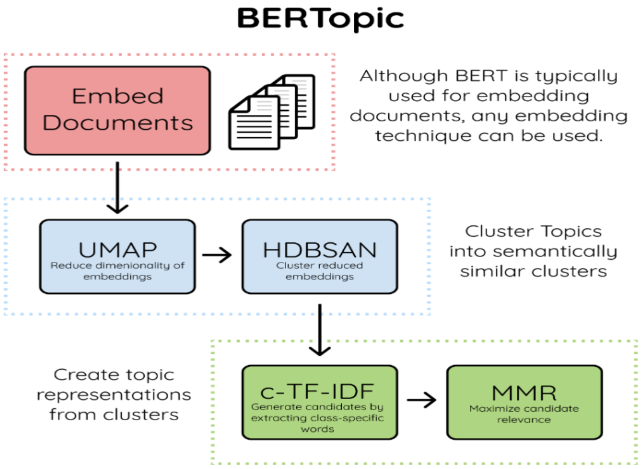


Other Interpretation visualization:





Internal Architecture of BERTopic:



BERTopic
Embed Documents
Although BERT is typically used for embedding documents, any embedding technique can be used.
UMAP
Reduce dimenionality of embeddings
HDBSAN
Cluster reduced embeddings
Cluster Topics into semantically similar clusters
Create topic representations from clusters
c-TF-IDF
Generate candidates by extracting class-specific words
MMR
Maximize candidate relevance

The algorithm has three main components:

- **Embed Documents:** Creates Document embeddings from a set of documents using pre-trained sentence-transformers models.
- **Cluster Documents:** Uses UMAP to lower the dimensions by still preserving the local structure and then uses HDBSCAN to cluster similar documents.
- **Create Topic Representation:** Documents in the single cluster are treated as a single document and TF-IDF is applied. This extracts the most important words per cluster, which gives us the overall idea about that topic. This model is called class-based TF-IDF
- **Topic Coherence:** In practice, some words will overfit the documents like the signature of a person. To improve the coherence Maximal Marginal Relevance was used to find the most coherent words without having too much overlap in between the words. This technique removes the words that do not contribute to the topic,

LDA vs BERTopic:

Latent Dirichlet Allocation (LDA) gives the list of [num_topics x probability] that shows the probability that the document belongs to a particular topic. LDA model uses a vector for each document and applies methods. Since LDA is a very low dimensional, dimension reduction methods will not perform well. There are no topic modeling LDA vectors, it just gives the probabilities of a document belonging to a topic.

On the other hand, BERT is very simple to implement. Initially, it creates an embedding from these documents and uses those embedding as a source to other clustering algorithms. Umap and T-sne perform very well in BERT which indicated it has rich embeddings. LDA is the default method for topic modeling, but if we have context in our documents, and we are trying to make use of these sentence embedding BERT gives us better results.

Bibliography

1. Groontendorst, Maarten. 2020. "Topic Modeling with BERT." Towards Data Science. https://towardsdatascience.com/topic-modeling-with-bert-779f7db187e6.
2. Groontendorst, Maarten. 2021. "Interactive Topic Modeling with BERTopic." Towards Data Science. https://towardsdatascience.com/interactive-topic-modeling-with-bertopic-1ea55e7d73d8.
3. Liu, Lin, and Lin Tang. 2016. "An overview of topic modeling and its current applications in bioinformatics." *SpringerPlus 5, 1608 (2016).* https://doi.org/10.1186/s40064-016-3252-8.
4. Marteen. n.d. "BERTopic." https://maartengr.github.io/BERTopic/tutorial/algorithm/algorithm.html.