

Assignment 2 – ICT202 - 34732205

In June 2022, we reached a user base of 4.76 billion on social media [1]. While social media platforms play an important role in providing information, they can also be a source of outdated or misleading information. During 2020, the COVID crisis was widely followed on Twitter, leading to the propagation of false claims, particularly regarding the effects of vaccines. This scenario highlights the need for tools like Natural Language Processing (NLP) to identify and prevent the spread of invalid and inappropriate information.

The objective of this assignment is to analyze the trending topics on Twitter during the COVID crisis. The analysis will be performed using BERTopic to identify the topics within a dataset of 20,000 tweets. The assignment will cover every step, including data collection, pre-processing, selection, training, evaluation and interpretation of the results obtained.

I- Data nature and collection

The dataset used in this project has been collected from *Kaggle* [2]. The tweets of this dataset contain the #CovidVaccine hashtag. Since our focus is on creating an unsupervised learning model, we will only use the raw text of each tweet.

The #CovidVaccine hashtag included in all tweets provides a diverse range of content within the tweets. This diversity is ideal for conducting topic modeling as it allows various aspects related to the COVID vaccine to be analyzed.

Because the datasets provide all tweets from 2020 to the present, a subsample of the very first tweets has been selected.

For dataset reading, the **read_csv** function from the *pandas* library has been used.

II – Pre-Processing

As the pre-processing requirements differ depending on the data type, I chose to implement a custom pre-processing approach.

The initial step involved removing English **stop words** and **punctuation** from all sentences. The *nltk* library has been used to perform this step.

The next step involved applying a **stemming algorithm** using the *nltk* library. Stemming is a process that reduces words to their base form, and so eliminating variations of words. In our case, PorterStemmer was used to perform the stemming operation.

Finally, irrelevant components of the tweets were removed in the last step. To perform this last step, a tweet has been randomly chosen and the components have been analyzed to determine which information is irrelevant.



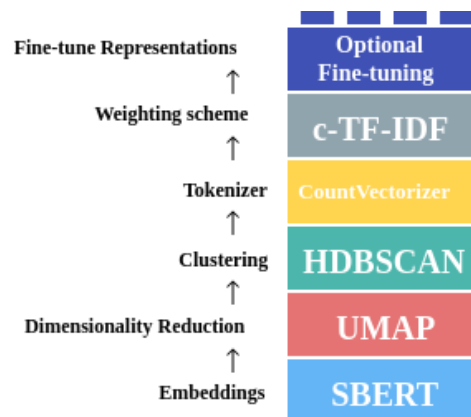
It appeared that **tags** and **url** were useless and so were removed in the pre-processing step. The **emojis** were also removed as they could not be treated as text.

III- Model selection and training

The BERTopic algorithm has been selected since it provides functions to interpret the results and evaluate the performance of the model.

Instead of using the bag-of-words approach which does not take the word's order into account, the BERT approach takes advantage of the contextual information given by the vector representation of BERT.

The model implementation in this assignment follows the convention used by BERTopic.



[3]

A) Embedding Model: SentenceTransformer

The goal of this step is to convert words into vectors to make it easier for manipulation and to reduce the time complexity. A **SentenceTransformer** has been used as it is a pre-trained library with a lot of different models.

The most used models have been selected: **roberta-base-nli-means-token**, **distilbert-base-nli-mean-token**, **stsb-roberta-base**, **paraphrase-distilroberta-base-v1**, **all-MiniLM-L6-v2**

A good embedding model provides similar vectors for similar words, a good context understanding and limits the size of the vectors generated.

The selected model was all-MiniLM-L6-v2 since it gives almost the same context understanding as the others with a vector size of only 384 against around 700-800 for the others.

B) Dimensionality Reduction Techniques: UMAP

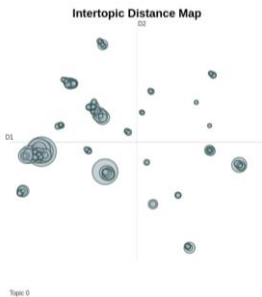

The second step involves performing dimensionality reduction to reduce the dimension of the input vectors (384x20000) in order to improve the performance of our model.

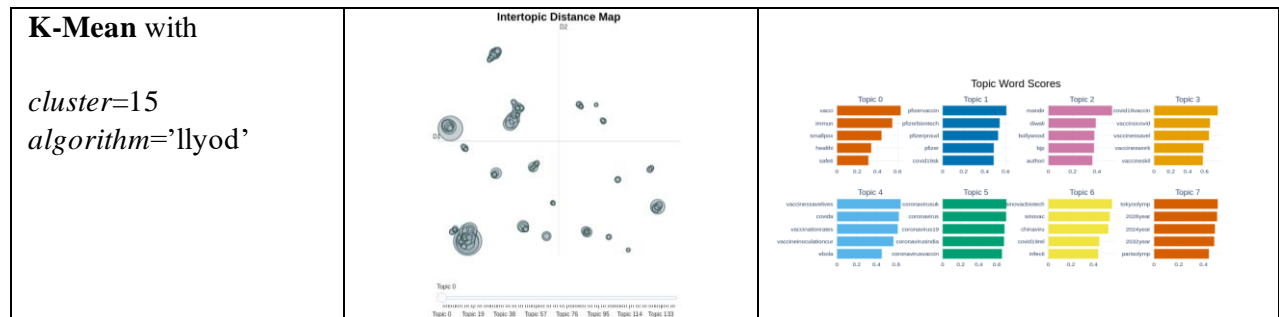
Two popular models have been selected: **PCA** and **UMAP**. While **PCA** significantly reduced the processing time, **UMAP** generated more accurate topics. This is because **UMAP** is more suited for capturing non-linear relationships in the data, which is particularly important in the context of tweet analysis.

C) Clustering embedding: HDBSCAN

The third step involves clustering similar vectors together to extract topics. This step is of utmost importance as the accuracy of the clustering model directly impacts the quality of the extracted topics.

A comparison between the two models (**HDBSCAN** and **K-Means**) and their respective hyperparameters such as *cluster size* and *metrics* has been made. Both **HDBSCAN** and **K-Means** got similar results in terms of topic extraction. However, **HDBSCAN** was a bit faster than **K-Means**, which makes it a better option for bigger datasets.

Clustering model specificity	Intertopic Distance Map	Topics
HDBSCAN with <i>min_cluster_size=15</i> <i>metric='euclidean'</i> <i>cluster_selection='com'</i> <i>gen_min_span_tree=True</i>		



D) Vectorizers: CountVectorizer

Once the topics have been grouped together, the next step is to generate relevant topics based on these clusters. In the BERTopic framework, the *CountVectorizer* serves as a token counter which allows the generation of topics.

For this step, a simple *CountVectorizer* has been used with a filter on the stop words.

E) Class-based TF-IDF Representation

The last step consists of weighing words in each cluster based on their frequency in the corpus. For this step, a simple C-TF-IDF model has been used.

IV) Topic model evaluation

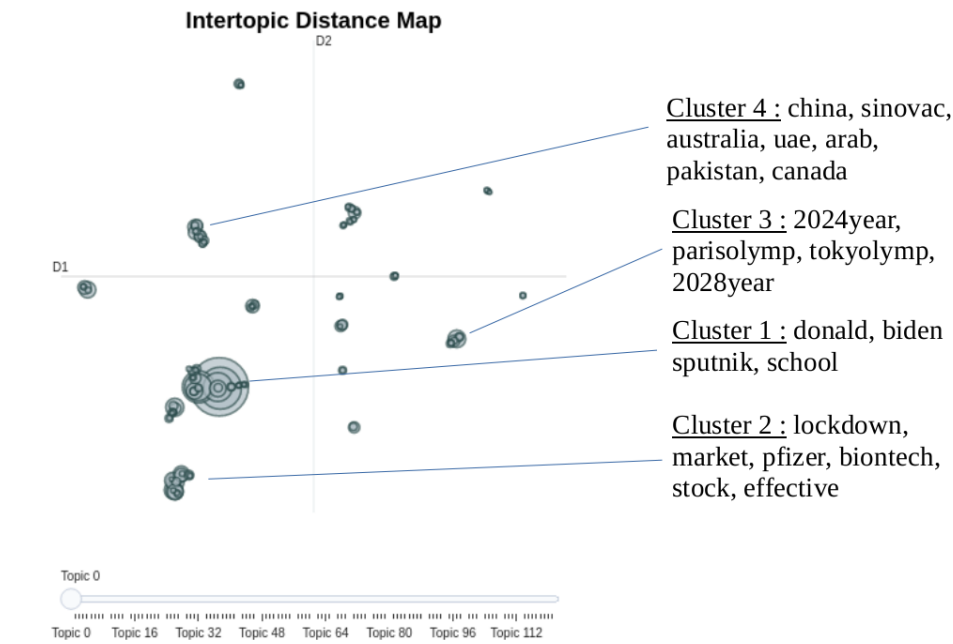
After selecting the model, an evaluation of the model has been made. We cannot use the usual metrics for supervised learning such as the F1 score. When it comes to evaluating an unsupervised topic model, there are 2 main metrics which are: topic coherence and topic diversity.

A) Topic coherence

Topic coherence is a measure of the interpretability of the generated topics. We can use a coherence model to measure the coherence. The coherence model provided a coherence score of around 0.41 for 20,000 tweets and around 0.50 for 4,000 tweets which is acceptable.

B) Topic diversity

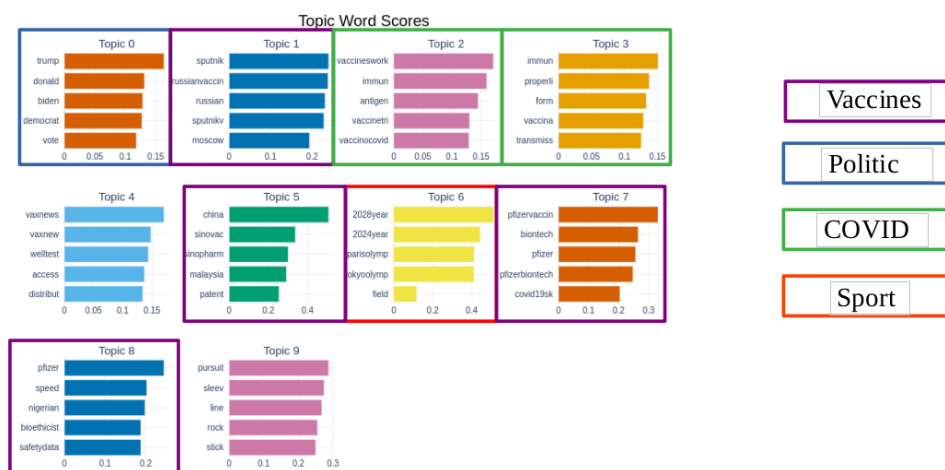
Topic diversity ensures that the generated topics cover a wide range of distinct concepts. To evaluate the topic diversity of the model, the intertopic distance map provided by BERTopic was analyzed.



What can be inferred from that graph is that each cluster provides different information about the same subject. Cluster 1 focuses on politics and education, while cluster 3 is about sports, and cluster 2 about the economic consequences. However, cluster 4 appears to lack coherence, which may explain the final coherence score.

V) Topic model results and interpretation

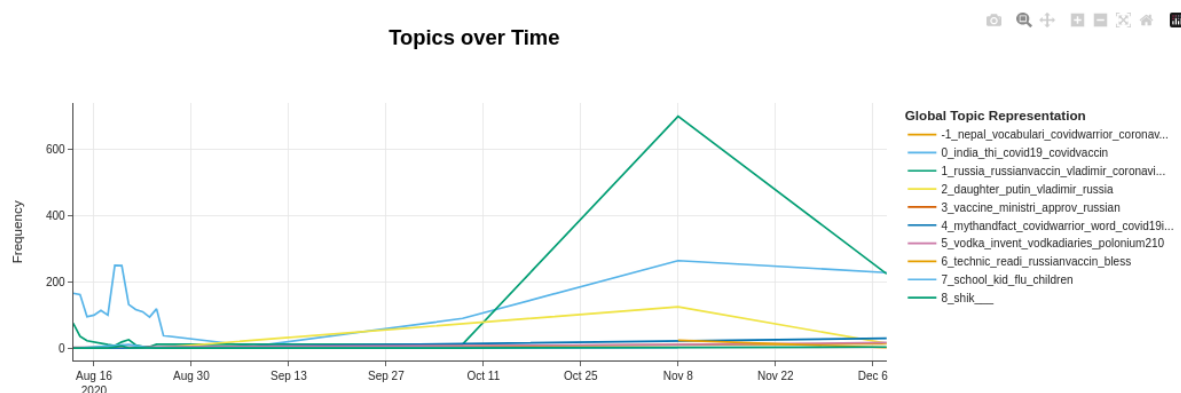
The topics and their associated words have been generated using the *visualize_barchart* method.



All topics are somehow related to the COVID vaccine. While certain topics directly address the vaccine such as topic 1, 5 and 8, topic 0 focuses on politics and topic 6 is about the Olympic Games.

With this model, we can now categorize tweet contexts and understand their relation to the COVID vaccine. This can be really useful when tracking false information.

To avoid the spread of misinformation, we need a model that monitors temporal trends. The following graph shows the topics trends between August 2020 and December 2020 (around 4,000 tweets).



By using this graph, we can analyze the trends and stop false information before it has a chance to circulate widely.

Conclusion

BERTopic has proven to be a valuable tool providing effective techniques. With BERTopic, we can track false information by identifying the trending topics, which is particularly crucial in critical contexts such as the COVID crisis.

What makes BERTopic even more interesting is its ability to monitor these trends daily by using tweet dates. This allows us to stay informed about the ongoing spread of fake information.

References

- [1]: Stats on social media, <https://www.statista.com/topics/1164/social-networks/#topicOverview>
- [2]: Dataset, <https://www.kaggle.com/datasets/kaushiksuresh147/covidvaccine-tweets>
- [3]: BERTopic algorithm, <https://maartengr.github.io/BERTopic/algorithm/algorithm.html>
- [4]: Tweet from the dataset, <https://twitter.com/NWDesignHub/status/1295631522741932032>