

Topic modelling using Natural Language Processing

2023 Summer Internship (20th May - 15th July)
Sidhant Guha (3rd Year CSE)
Kalinga Institute Of Industrial Technologies

Guide Name: Prof. Arindam Mondal & Prof. Asimabha Bhowmick

Context



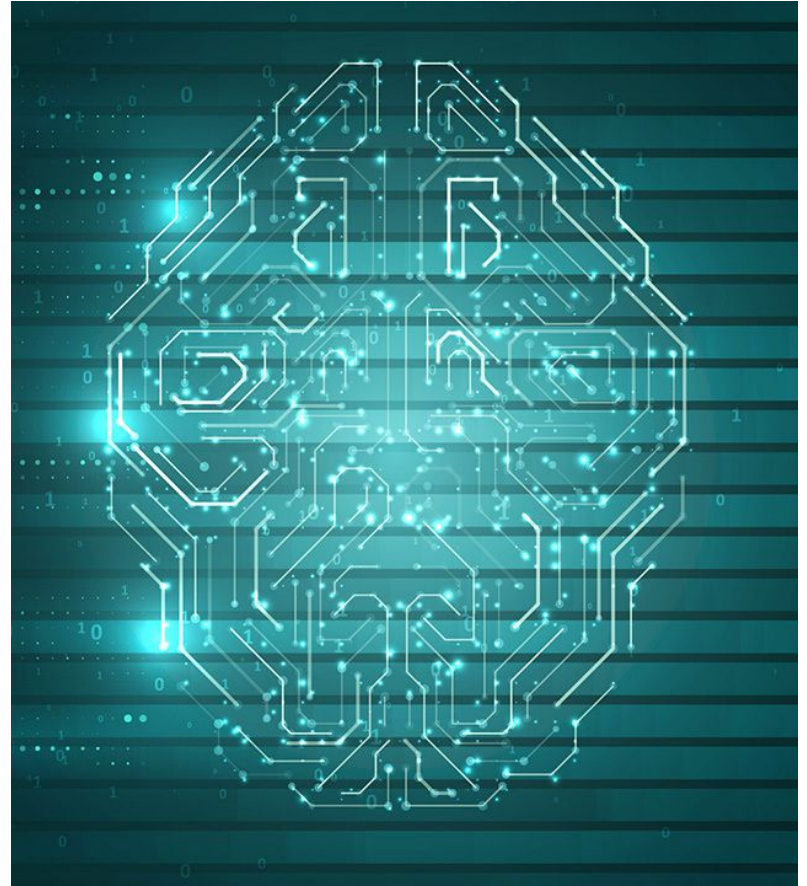
As part of research work at XLRI Jamshedpur India, Prof. Arindam Mondal and Mr Asimabha Bhowmick are working on a family business review system to understand the evolution of family businesses over the past few decades. Their study aims to provide an understanding of the kind of research work that has informed scholars about this area.

As part of my internship assignment, I was tasked to employ a method of topic modelling which consisted of understanding how natural language processing can be applied to a very large dataset for analysis and identify the most important aspects of the documents. The idea was to use the latest NLP model developed by Google known as BERTopic.


Objective

The objective of this project was to implement BERTopic on a sample dataset containing a list of different articles and its details. The dataset required thorough study using BERT so as to extract the underlying topics (*latent themes) from the documents. The next step was to make it as efficient as possible using BERT and visualize it using different methods like graphs, heatmap etc.

** **Latent Themes:** Semantic codes and themes identify the explicit and surface meanings of the data. The researcher does not look beyond what the participant said or wrote. Conversely, latent codes or themes capture underlying ideas, patterns, and assumptions. This requires a more interpretative and conceptual orientation to the data.*



What is Natural Language Processing?



Natural language processing is a part of artificial intelligence which specializes on enabling the computer to comprehend human language. Through this it is able to generate answers to questions and manipulate them depending on what the user wants. For instance NLP is the core technology behind virtual assistants like Siri, Cortana and Alexa by which they are able to interpret the human questions.

Applications of NLP:

- Chatbots powered by NLP can process a large number of routine tasks that are handled by human agents today, freeing up employees to work on more challenging and interesting tasks
- Advancements in NLP improves the computer's understanding on what the user actually wants to know
- NLP can analyze customer reviews and social media comments to make better sense of huge volumes of information

What is BERT?

Bidirectional Encoder Representations from Transformers (BERT) is a family of language models introduced in 2018 by researchers at Google. The BERT framework was pre-trained using a large source of text from Wikipedia and can be fine-tuned with question and answer datasets. We can then apply the training results to other Natural Language Processing (NLP) tasks, such as question answering and sentiment analysis.

B E R T

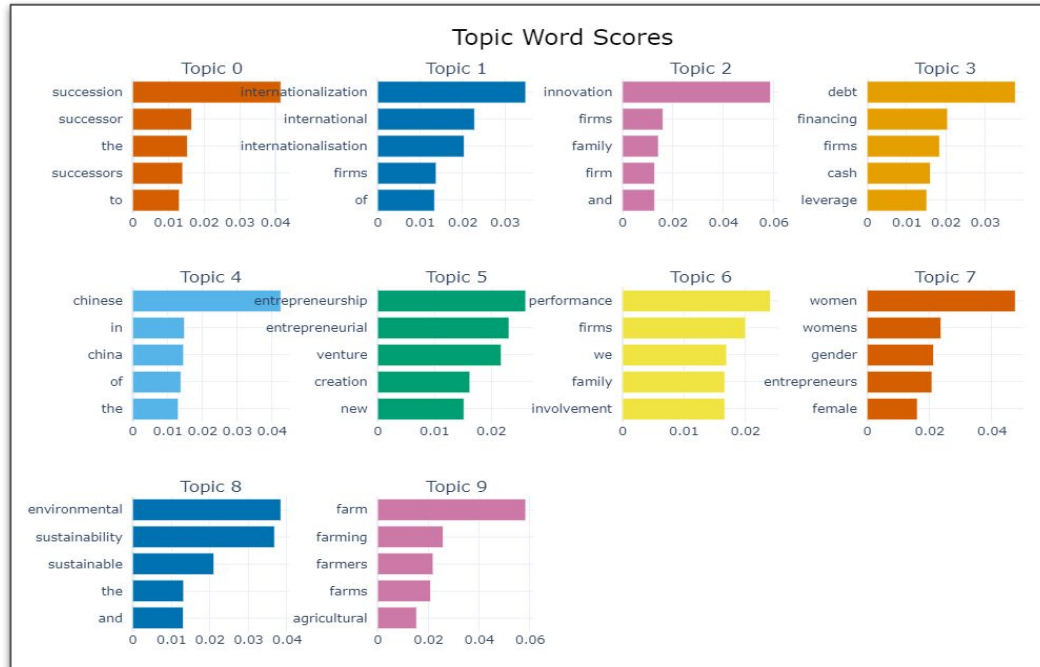
As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once. Therefore it is considered bidirectional, though it would be more accurate to say that it's non-directional. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word).

Encoders in Transformers are machines that process the input sequence (collection of words) and produce a continuous representation, or embedding, of the input that the computer is able to understand.

A transformer is a model that is self-sufficient and evaluates input and output data representations.

Transformers are primarily applied in computer vision and Natural Language Processing. They are also used in machine language translation, conversational chatbots, and search engines.

Bar chart for top N(10) topics



These bar charts show which words are most frequently occurring in the top n topics (in this case n=10).

By referring to these scores we get to know what the topic is about.

Analyzing the Data



I was provided with a sample corpus on which I trained the BERTopic model and further made it more efficient by clustering the data and using the techniques which BERT provides like UMAP and HDBSCAN. Then I implemented the code to visualize the topics like displaying the word clouds of the top n topics and included features like graphs, heatmap, matrix etc.

UMAP, which stands for Uniform Manifold Approximation and Projection, is a dimensionality reduction technique used for visualizing high-dimensional data in a lower-dimensional space. Dimensionality reduction is the task of reducing the number of features in a dataset.

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that is useful for discovering clusters of varying densities in your data. Clustering is the process of dividing similar data into a number of groups such that the data values in that group are more similar to other data points in that same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

What is High-dimensional data?

When the number of features in a dataset exceeds the number of observations it is known as a high dimensional data. For example : Healthcare datasets where the number of features for a given individual can be massive (i.e. blood pressure, resting heart rate, immune system status, surgery history, height, weight, existing conditions, etc.).

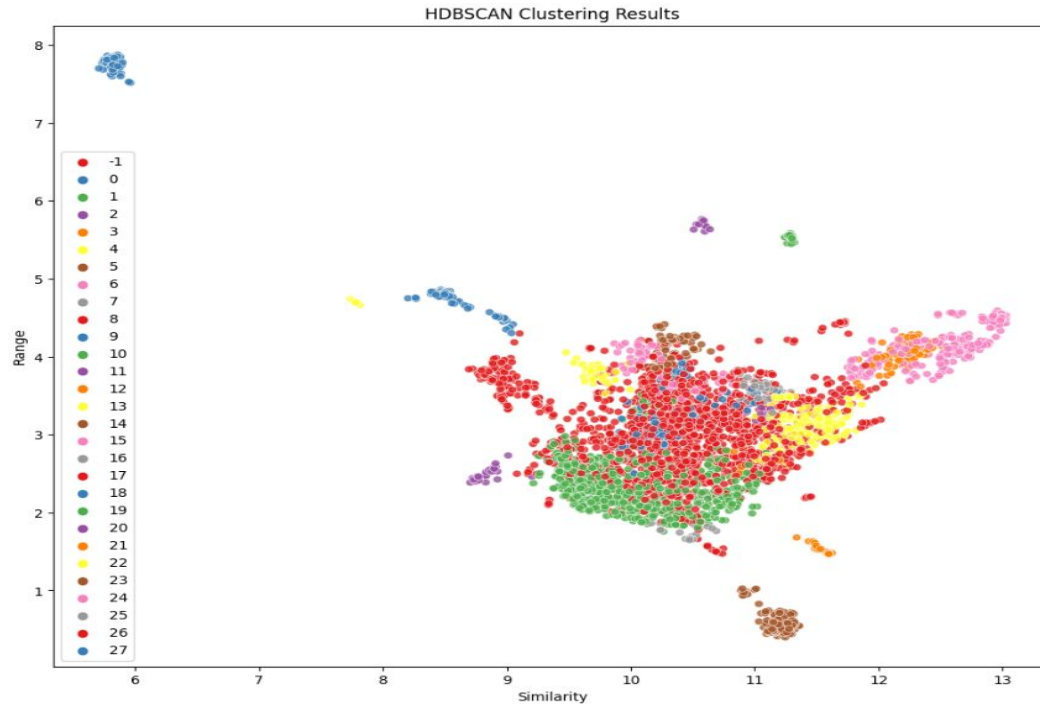


The diagram shows a table with 7 columns and 4 rows. The first column is empty. The next six columns are labeled 'Blood pressure', 'Heart rate', 'height', 'weight', '...', and '...'. A red bracket above the last five columns is labeled 'features'. A red bracket to the left of the last three rows is labeled 'observations'.

	Blood pressure	Heart rate	height	weight
Person 1						
Person 2						
Person 3						

Therefore Low-dimensional data is the opposite of High-dimensional data that is the number of features in the dataset are either equal or less than the number of observations.

HDBSCAN Visualization



This is the graphical representation of the HDBSCAN algorithm.

This graph shows us the similarity between the topics(clusters) and how much dense they are that is which topics are important and have the most contribution in the database. For example Topic 1,10,19(Green) and Topic 8,17,26(Red) are similar and have high importance in the database.

Word Cloud

By this word cloud we can make a probable assumption on what the cluster is about. Like for example cluster 14 is more about the role of women in family businesses and how they have different position in the society as displayed below.

Cluster 14 Word Cloud



Outcome



Through this internship, I was able to provide a prototype on how BERTopic works and how it is highly efficient in dimensionality reduction and density based clustering along with visualizing different topics to help the researcher understand and analyze the Family Businesses over the years. I employed the novel method BERTopic of machine learning to identify research topics, which has the advantage that no prior knowledge of the topics is required.

The structure of BERTopic (embeddings, UMAP, HDBSCAN), allows for a very flexible algorithm that can easily adapt to new advancements in language models, clustering techniques, dimensionality reduction techniques, etc. Thus far, every time a new sentence-transformer has been released, the resulting quality of topics has been increased (at least in my opinion). The same should apply to the other models.

This minimizes necessary judgment by the researcher and hence increases objectivity. However, judgment is still required to interpret the final topics.

Concluding Remarks



This internship adds a great value to my experience and how I was able to learn new things on the go. I will delve more into the field of ML/AI and learn more about it. Using this experience I should be able to crack more internships in the future where I can further improve my skills.

I intend to keep working with Asimabha Sir for a few more days to further improve my work so as to help him as much as possible with his research. Now I am currently working on how we can make the topic model more efficient using both BERT and LDA.

References

- Getting started with the built-in BERT algorithm :
<https://cloud.google.com/ai-platform/training/docs/algorithms/bert-start>
- Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing by Google Blog:<https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html>
- BERTopic for Topic Modeling - Maarten Grootendorst - Talking Language AI:https://www.youtube.com/watch?v=uZxQz87lb84&t=761s&ab_channel=Cohere
- BERTopic Explained by James Briggs:https://www.youtube.com/watch?v=fb7LENb9eag&t=1178s&ab_channel=JamesBriggs
- How to use BERTopic - Machine Learning Assisted Topic Modeling in Python:https://www.youtube.com/watch?v=v3SePt3fr9g&t=70s&ab_channel=PythonTutorialsforDigitalHumanities
- BERT Transformers for Language - EXPLAINED!:https://www.youtube.com/watch?v=TLPmlVeEf1k&t=1730s&ab_channel=CodeEmporium
- BERT Explained: State of the art language model for NLP:<https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>
- Understanding UMAP:<https://pair-code.github.io/understanding-umap/>
- How HDBSCAN works?:https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

Thanks

In this project I was helped by
Dr.Arindam Mondal, Mr.Asimabha
Bhowmick, Mr.Maarten Grootendorst
and Mr.James Briggs.

