**Abstract**

This paper regards the result and analysis of the capstone project, where BERTopic was used alongside with Llama2, kinda like an ensemble of the two. The difference between a topic model and a large language model is seen clearly in this results/analysis paper. This paper should provide a comprehensive application for topic modeling and an assistant to research especially in the area of Machine learning Arvix published papers.

**Introduction**

In the recent event, the field of machine learning and Natural language processing has been significant advancing. This project aims to leverage these advancements by developing a comprehensive application that utilizes ML models for analyzing machine learning research papers. The core of the application is the BERTopic model integrated with the infamous Llama 2 13B model by Meta, using Huggingface for enhanced text generation capabilities.

**Methodology**

The project employs a variety of libraries and tools to achieve the objectives.

Libraries Used: ‘bertopic’, ‘datasets’, ‘accelerate’, ‘bitsandbytes’, ‘xformers’, ‘adjustText’, and ‘textract’. ‘Streamlit’ for future works. For dataset: The ML-ArXiv-Papers dataset from the ‘datasets’ library is used which includes the abstracts and titles of machine learning papers. The model Llama-2-13b-chat-hf is configured with the 4-bit quantization to reduce the GPU memory usage, though the model runs on a GPU if available better.

The BERTopic model is used with sub-model for embedding (SentenceTransformer), dimensionality reduction (‘UMAP’), clustering (‘HDBSCAN’) and topic representation (including ‘KeyBERTInspired’, (KeyBERT) ‘MaximalMarginalRelevance’ (MMR) and ‘TextGeneration’ using Llama2.

**Findings / Results:**

The application successfully generates topics from the dataset of machine learning research papers. Key findings include:

The use 4-bit quantization effectively reducing the memory usage of the Llama2 model allowing to run on GPUs with limited memory. The integration of various sub model provided a thorough approach for topic modelling that also revealed the underlying themes in the dataset. The use of UMAP for dimensionality reduction and the visualization of documents and topics enables interactive exploration of the model's output leading to a quick oversight of the dataset.

The highlight, Llama 2 for text generation can be considered a bit of a novel approach in topic representation and labeling. This approach goes a long way when C-level executives are taken into consideration for the usage of the same.

An example of the same is shown in the screenshot at the end of this paper. I would like for you to use the notebook and experiment with the chatbot (very basic but efficient).

**Conclusion**

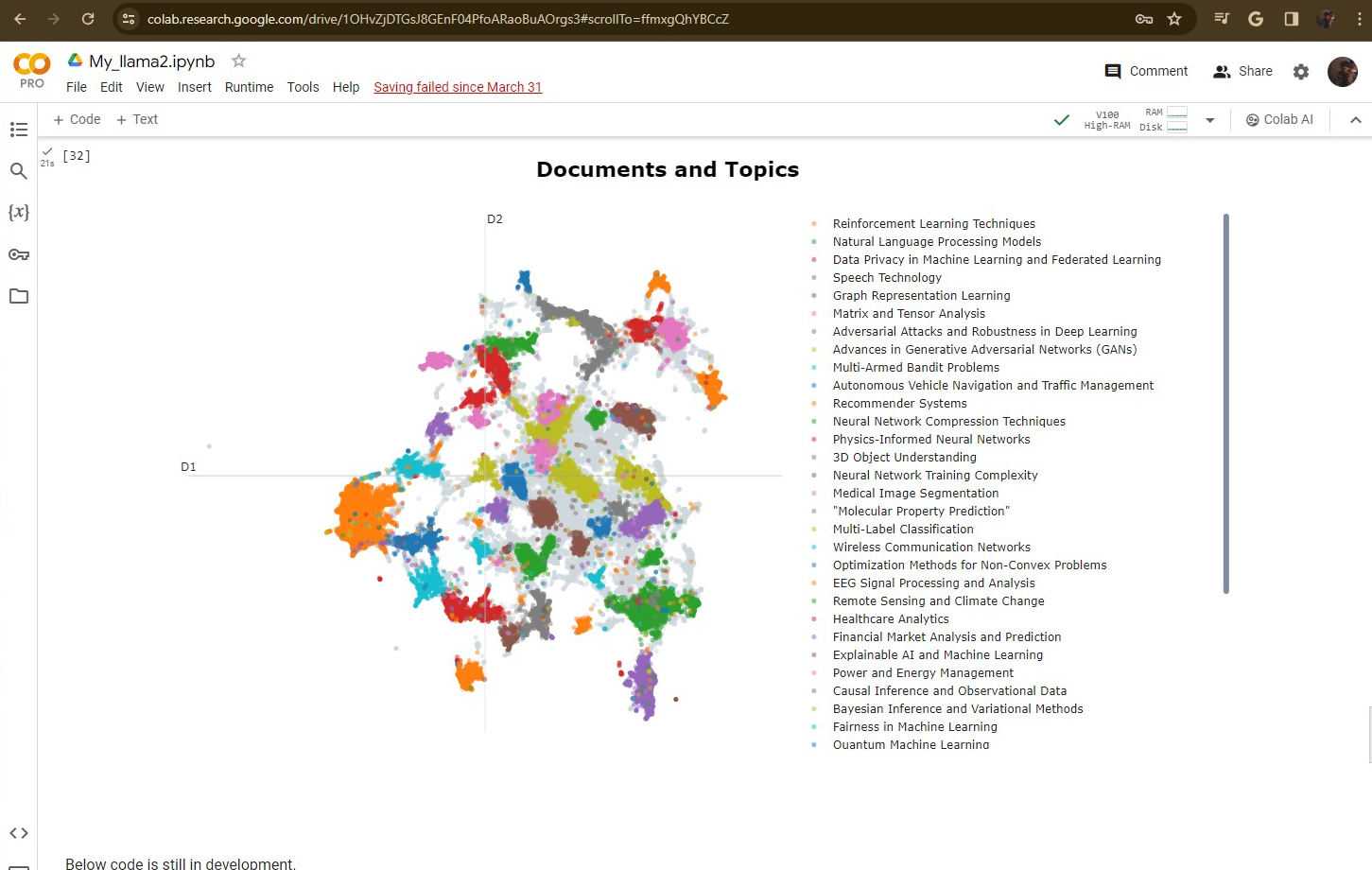
The project has the potential of combining advanced NLP techniques and models for topic modeling and analysis. The approach of using llama 2 for text generation in topic modelling is particularly an game changer. However, this comes at a cost and the challenges in terms of resource optimization harm the further development to enhance the application’s capabilities. This is a step forward in applying the NLP techniques for topic modeling and analysis, showing insights into the content and context of the dataset.

**Future work**

There is definitely a need to develop the chatbot to be appealing and easier to use, while implementing the ghosting mechanism in Llama 2, which can further help users like researcher and ML developers to be fixed on the topic they are researching or solving. Continued efforts to optimize the code for resource usage, especially when working with large models and datasets is a must. If this can be done then there can be an increase in usability of this model by a variety of audience.

A screenshot of a computer

Description automatically generatedScreenshots from the notebook and link for the same is [here](https://colab.research.google.com/drive/1OHvZjDTGsJ8GEnF04PfoARaoBuAOrgs3?usp=sharing). (It is still private but if you request access then I will make sure to be prompt with allowing access to you only.)



The number clusters were created by a choice, a minimum of 500 abstracts in a cluster. Something that can be worked on or changed by choice and further review. Below is the topic differentiation by Model.

A screenshot of a computer

Description automatically generated