# Unsupervised Abstractive Opinion Summarization with BERT

## Description:

Sentiment analysis is the computational study of people's appraisals and emotions toward entities, events, and their attributes. The goal is to determine whether an opinionated document (e.g., reviews) or sentence expresses a positive or negative opinion. Opinions are essential because whenever someone wants to make a decision, they would like to hear others' opinions [1].

In parallel, many works have been done in multi-document summarization given a collection of documents, produce a concise summary. Most of the existing works generate extractive summaries, where the output is copied from the input texts. Another paradigm, abstractive summarization - where the summary is generated word by word – is extensively use for single-document summarization, thanks for large datasets.

However, 1) opinions or sentiments are not explicitly taken into account 2) there is a lack of large datasets in multi-document summarization because the annotation is extremely costly and does not scale. Therefore, we have a lack of structure in the former, and in the latter, only extractive models can be developed.

Recently, there is a new trend of unsupervised abstractive neural multi-document summarization. [2] was the first to propose an autoencoder-like architecture to 1) encode-decode reviews separately, 2) aggregate them using the mean operator as a "summary" embedding, and then decode it. They show the model was able to general summaries, although not perfect. In the last months, other works have been going into this direction with slight variations: [3,4,5,6,7,8].

In this master project, we would like to go beyond and generate more structured abstractive summaries in an unsupervised way. We would use a pipeline producing large datasets and develop our model(s) on it. Moreover, we would leverage recent large pre-trained language models such as BERT [9]. Ideally, we would publish the work at NAACL 21[1].

## Prerequisites:

- Knowledge about NLP and Machine Learning. Experienced with Python and PyTorch.
- Understand most of the previous work [2-8].
- Being motivated and ready to produce publishable work.
- (Optional) experienced with *HuggingFace Transformers* library[2] .

**Supervisor**: Diego Antognini (diego.antognini@epfl.ch)

## References:

[1] https://www.cs.uic.edu/~liub/FBS/IEEE-Intell-Sentiment-Analysis.pdf
[2] https://arxiv.org/pdf/1810.05739.pdf
[3] https://arxiv.org/pdf/2004.14884.pdf
[4] https://arxiv.org/pdf/1911.02247.pdf
[5] https://arxiv.org/pdf/2004.10150.pdf
[6] https://arxiv.org/pdf/2004.14754.pdf
[7] https://arxiv.org/pdf/2005.01901.pdf
[8] https://arxiv.org/pdf/2006.00119.pdf
[9] https://arxiv.org/pdf/1810.04805.pdf

[1] https://naacl.org/ deadline around December.

[2] https://github.com/huggingface/transformers .