Sentiment analysis

In this project, you will experience how to **build a sentence-level sentiment analysis model** from scratch with real-world data.

1. Collect the data from a variety of sources
   - Choose a target product of interest and collect corresponding reviews from the seller's web pages and any related pages
   - You may consider Selenium for crawling web pages and BeautifulSoup for parsing the pages.
   - The text can be in either English or Vietnamese. Handling Vietnamese text is more challenging, but it may improve your score.
2. Preprocess the collected data and label the sentences
   - Break each of the above reviews into multiple sentences: 1) each sentence is a simple sentence, and 2) the content of the sentence subjects to the target product.
   - Normalize the sentence heuristically, e.g., remove spelling mistakes and typos, fix the informal abbreviations, etc.
   - Manually label each sentence as Positive or Negative. Please avoid ambigous sentences.
   - Your dataset after processing should include at least **3000 positive** and **3000 negative** samples. However, the more data, the better model.
3. Create the embeddings for each of the above sentences using the following text representations: **TF-IDF**, **fastText**, and **BERT** (or PhoBert for Vietnamese)
4. Build your sentiment analysis model using any conventional machine learning approach and report the **accuracy**, **precision**, **recall**, and **F1-score** overall, as well as for each class.

You need to prepare the following materials for your group submission.

- The link(s) to the Google Colab implementation. You need to organize the datasets such that the grader can run your implementation with minor adjustments. You also need to guarantee that no edit is made after the deadline.
- A document to present how you accomplish the tasks in Questions 1 to 4

Important notes:

- This project gives you a 20% course grade.
- Strictly avoid plagiarism in any circumstance.