HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY
FACULTY OF COMPUTER SCIENCE

IT4142E

SEMESTER 2023.1

MOVIE REVIEW SENTIMENT ANALYSIS

INTRODUCTION TO DATA SCIENCE

CLASS CODE: 144108

OUR GROUP INCLUDES FIVE MEMBERS:

Tạ Ngọc Minh	20214918	minh.tn214918@sis.hust.edu.vn
Lê Ngọc Bình	20214878	binh.ln214878@sis.hust.edu.vn
Trần Thanh Trường	20214938	truong.tt214938@sis.hust.edu.vn
Hoàng Đình Dũng	20214882	dung.hd214882@sis.hust.edu.vn
Ngô Việt Anh	20214875	anh.nv214875@sis.hust.edu.vn

SUPERVISOR: Associate Professor Thân Quang Khoát

Hanoi - 2023

Sentiment Analysis: *Aspect-Based Sentiment Analysis using film review*

[*]Project Report of the course IT4142E - Introduction to Data Science

Ta Ngoc Minh[*] — *SID No. 20214918* — *Email: minh.tn214918*[†] — HUST-SoICT

Le Ngoc Binh — *SID No. 20214878* — *Email: binh.ln214878*[†] — HUST-SoICT

Tran Thanh Truong — *SID No. 20214938* — *Email: truong.tt214938*[†] — HUST-SoICT

Hoang Dinh Dung — *SID No. 20214882* — *Email: dung.hd214882*[†] — HUST-SoICT

Ngo Viet Anh — *SID No. 20214875* — *Email: anh.nv214875*[†] — HUST-SoICT

***Abstract*—This study introduces a novel method for sentiment analysis in the film industry, using Aspect-Based Sentiment Analysis (ABSA) enhanced by Large Language Models (LLMs). It focuses on detailed examination and measurement of sentiments in movie reviews, specifically targeting distinct aspects such as acting, direction, cinematography, screenplay, and special effects. The approach involves advanced NLP techniques and combines supervised and unsupervised learning, leveraging the deep contextual understanding of LLMs. The research uses a vast collection of film reviews across various genres and styles for a thorough analysis.**

The goal is to provide deeper insights into public opinion, aiding filmmakers, critics, and marketers by accurately evaluating sentiments linked to specific film aspects. This can improve movie recommendation systems and offer a precise tool for assessing audience reactions. Additionally, the study contributes to the fields of NLP and sentiment analysis by showing how LLMs can be effectively applied in specialized areas. It enhances the understanding of the relationship between cinema and its audience and has the potential to transform sentiment analysis in the film industry.

I. Introduction

THE REALM OF FILM CRITICISM AND ANALYSIS presents a rich tapestry of subjective opinions, diverse perspectives, and emotive expressions. Understanding these sentiments not only helps in gauging public opinion but also serves as a crucial tool for filmmakers, marketers, and audiences alike. Our project embarks on an ambitious journey to harness the power of Large Language Models (LLMs) in analyzing sentiments expressed in film reviews.

The primary focus of our project is to delve into the depths of film reviews, extracting nuanced sentiments that go beyond the traditional positive, negative, or neutral categorizations. Film reviews are a complex blend of emotions, critiques, and appraisals that require a sophisticated level of understanding, one that LLMs like GPT-3 and BERT are well-equipped to provide.

In this project, we leverage the advanced natural language processing capabilities of LLMs to dissect and analyze the sentiment of each review. The goal is to capture the multi-faceted nature of opinions expressed by reviewers, ranging from their thoughts on storytelling and direction to acting, cinematography, musical score, and special effects. Each of these aspects contributes to the overall sentiment of the review and, by extension, to the perceived success or failure of a film.

One of the unique challenges in this endeavor is the subjective nature of film reviews. Unlike product reviews, which often have clear-cut positive or negative sentiments, film reviews are a blend of personal taste, artistic interpretation, and emotional response. This complexity demands an LLM-based approach that can understand context, detect subtleties in language, and interpret the underlying tones of a reviewer's language.

Our project is not just about sentiment analysis; it's about understanding the language of cinema through the lens of its audience. We aim to build a system that can provide insightful analytics to filmmakers, critics, and movie enthusiasts, offering a deeper understanding of what resonates with audiences. This involves not only analyzing existing reviews but also predicting audience reactions to new releases based on historical sentiment trends.

Through this project, we aspire to create a bridge between the art of filmmaking and its audience, using the cutting-edge capabilities of Large Language Models. We believe that this endeavor will not only contribute significantly to the field of sentiment analysis but also offer valuable insights into the ever-evolving relationship between cinema and its viewers.

II. Data collection and preparation

A. Data selection

In conducting this study, especially when dealing with movie reviews sourced from online platforms, a cautious approach in dataset selection is vital to avoid biased algorithmic decision-making that could lead to unfair treatment of individuals based on race, color, or nationality. To ensure a balanced and fair representation of audience opinions, we have decided to utilize datasets from both IMDB and Rotten Tomatoes, each offering unique insights and value.

[†] @sis.hust.edu.vn ; [*]Team leader.

IMDB and Rotten Tomatoes employ different methodologies for their ratings. IMDB uses a weighted average based on votes from its registered users, a technique that helps to normalize the ratings by reducing the impact of extreme values in a potentially asymmetrical distribution. Rotten Tomatoes, on the other hand, calculates its Tomatometer score based on the percentage of critics giving positive reviews. This approach has been noted for its systematic bias, particularly in its correlation with box office performance.

The demographic breakdown provided by IMDB, which includes age and gender, though limited geographically to U.S and Non-U.S categories, offers valuable insights into the general audience's perspective. This is an area where Rotten Tomatoes falls short, as it lacks detailed demographic data. Additionally, Rotten Tomatoes has been observed to have a gender bias, with studies indicating that 91% of its top critics are male, compared to a slightly more balanced 70% male demographic on IMDB.

While IMDB reflects the views of the general audience, Rotten Tomatoes captures the opinions of a selected group of professional critics. The general audience typically looks for entertainment and engaging stories, whereas critics often have more specific and stringent criteria in their evaluations. By combining datasets from both IMDB and Rotten Tomatoes, our study aims to capture a comprehensive range of viewpoints, encompassing both the general audience and professional critics. This dual-dataset approach is expected to provide a more holistic understanding of movie reviews, influencing the decisions of a wide spectrum of moviegoers. Therefore, the integration of data from both IMDB and Rotten Tomatoes is deemed essential for the success of this study.

B. Web scraping

Figure 1. Web scraping from websites

In this project, we implement a strategic approach in data analysis by utilizing two distinct datasets: film details for Exploratory Data Analysis (EDA) and film reviews for training Large Language Models (LLMs). This bifurcated approach allows us to harness the specific strengths of each dataset for different aspects of our comprehensive study in the cinematic domain.

This section details the methodology employed to collect film review data from IMDB and Rotten Tomatoes using web scraping techniques. Web scraping, the process of extracting data from websites, is a crucial step in gathering the necessary information for our sentiment analysis study. For this purpose, we utilize two popular Python libraries: Selenium and BeautifulSoup.

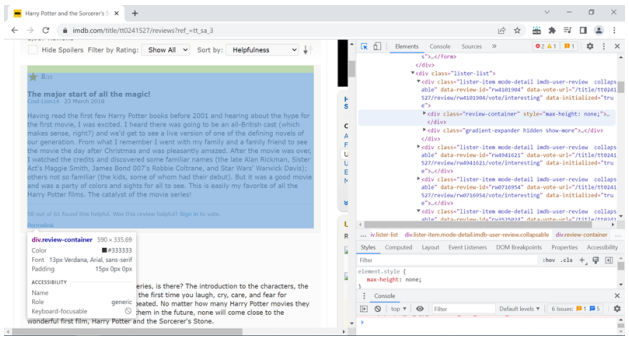

Figure 2. Scrape IMDB website

Selenium is an automation tool that is highly effective for navigating web pages that require user interaction or have dynamically loaded content. Since both IMDB and Rotten Tomatoes have interactive elements and potentially complex navigation systems, Selenium is ideal for accessing the data on these sites.

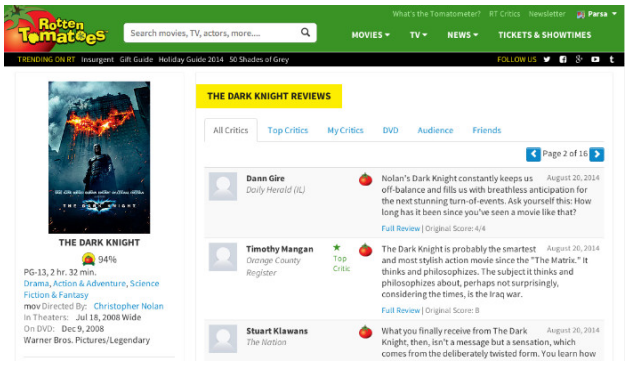

Figure 3. Scrape Rotten Tomatoes website

Once the required page is loaded with Selenium, BeautifulSoup is used for parsing HTML and extracting the needed data.

In addition to scraping film reviews, our methodology includes the extraction of detailed data about the films themselves from IMDB and Rotten Tomatoes. This comprehensive approach provides a more robust dataset, enabling a richer analysis that goes beyond sentiment analysis of reviews to encompass an understanding of how specific film attributes may influence audience and critic perceptions.

- **Film Details Data for EDA**

  - *Purpose and Goals*: The film details dataset is dedicated to EDA, aimed at uncovering trends, patterns, and correlations in the film industry. Crucial for understanding the influence of film attributes on ratings and popularity.
  - *Analysis Techniques*: Use of statistical methods and visualization techniques like histograms, scatter plots, and heatmaps to analyze the film industry's nature.
- **Film Reviews Data for LLM Training**
  - *Purpose and Goals*: Key for training LLMs to perform sentiment analysis on film reviews, providing a detailed understanding of public opinion.
  - *Model Training*: Involves preprocessing reviews and employing techniques such as tokenization and sentiment labeling. Aims to fine-tune a pre-existing LLM for enhanced sentiment analysis accuracy.
- **Ethical Considerations and Compliance**
  - **Adhering to Terms of Service**: Strict adherence to the terms of service of IMDB and Rotten Tomatoes during data scraping.
  - **Respectful Scraping Rate**: Implementation of a respectful scraping rate to avoid overloading servers.
  - **Secure Data Storage and Usage**: Secure storage and use of data exclusively for research purposes to uphold data integrity and privacy.
- **Challenges and Solutions in Web Scraping**
  - *Dynamic Content Loading*: Utilization of Selenium's wait functions to ensure full page load before data extraction.
  - *Changing Web Structures*: Regular updates to scraping scripts to accommodate website HTML structure changes.
  - *Data Quality*: Implementation of checks and validation to ensure data accuracy and completeness.

Web scraping with Selenium and BeautifulSoup offers a robust approach to collecting comprehensive and nuanced film review data from IMDB and Rotten Tomatoes. This data forms the backbone of our sentiment analysis study, providing the raw material needed to extract insights into public opinion on films.

C. Data preprocessing and Storing

It is crucial to keep the dimensionality of the text low to improve the performance of the model. Thus, it is highly recommended to remove the noise as much as possible and to properly preprocess the text in this pre-analysis stage.

- Contractions: Dictionary of contractions is provided to be filtered in each review and contractions will be expanded. All the capital letters will be converted to lowercase.
- Punctuations and special characters: and html special characters will also be removed.
- Tokenization: Each review will be separated into smaller units called tokens using NLTK package tokenization.
- Stopwords: will be removed using nltk package's built-in function. , POS tagging: A generic POS tagging is applied to classify words into four categories of adjective, verb, noun, and adverb.
- Lemmatization: The reviews will be lemmatized based on the POS tagging so it is crucial to have accurate tagging classification.

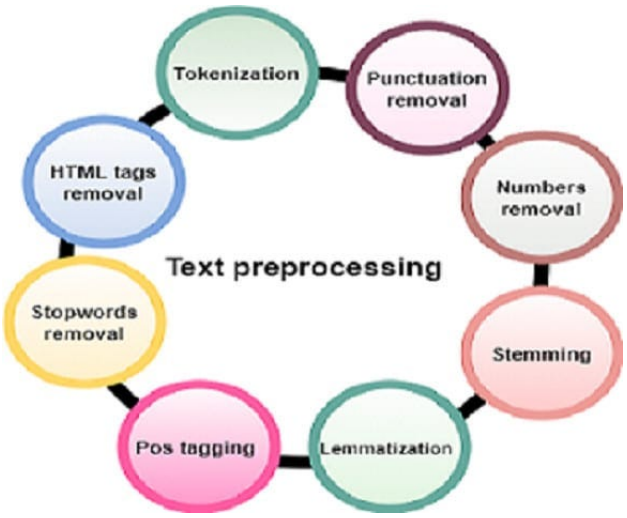

Figure 4. Text preprocessing

After the meticulous collection and processing of data, the next crucial step in our methodology involves the integration and conversion of this refined data into a structured format. Initially, we integrate the separately collected datasets — the detailed film data and the film reviews. This integration is carefully executed to ensure that each review is accurately associated with its corresponding film, allowing for a comprehensive dataset that combines both qualitative and quantitative aspects of the films.

Once integrated, the dataset is converted into a CSV (Comma-Separated Values) format. The choice of CSV is due to its wide compatibility with various data analysis tools and its ease of use in handling large datasets. This format also facilitates efficient data manipulation and querying in subsequent analytical processes.

Finally, the CSV file is securely stored in a database. The database serves as a centralized repository, ensuring organized storage and easy retrieval of data. It supports robust data management, allowing for efficient queries and scalable storage solutions. By storing the data in a database, we also enhance the security and integrity of the data, ensuring that it is accessible only to authorized personnel and is used strictly for the intended research purposes. This step marks the transition from data collection and processing to in-depth analysis and model training, setting the stage for insightful discoveries and advancements in our study.

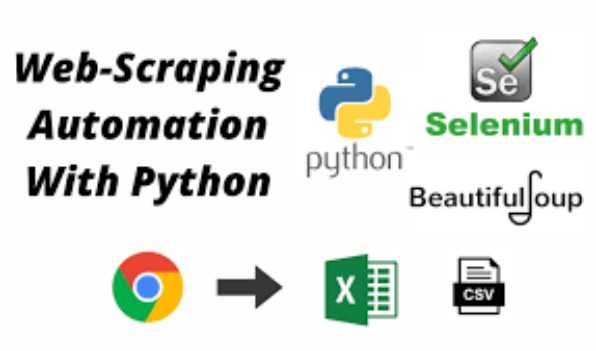

Figure 5. Data integrating

III. Exploratory Data Analysis

Exploratory Data Analysis. It is an approach to analyzing data sets to summarize their main characteristics, often with the help of statistical graphics and other data visualization methods. The primary goal of EDA is to uncover patterns, relationships, and insights in the data, which can then inform more targeted analyses.

We divided our EDA in 4 followings type:

- **Univariate Non-graphical:** Analyzing a single variable without visual representations to summarizing and describing the characteristics of a single variable.
- **Multivariate Non-graphical:** Analyzing the relationships between three or more variables without visual representations to examining relationships and dependencies among multiple variables through quantitative measures.
- **Univariate Graphical:** Analyzing a single variable using visual representations to visualizing the distribution and characteristics of a single variable.
- **Multivariate Graphical:** Analyzing the relationships between three or more variables using visual representations to visualizing complex relationships and patterns among multiple variables.

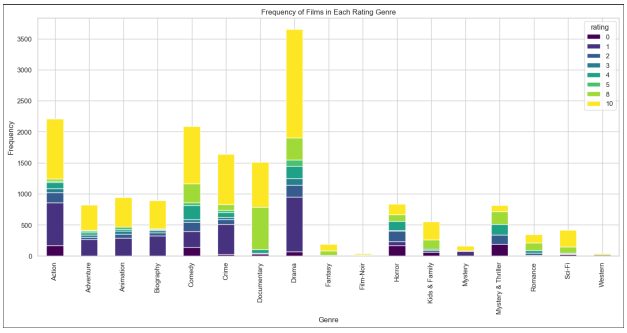

Figure 6. Stacked bar-chart for relationship between Ratings and Movie Genre

The above stacked bar-chart represent the distribution of rating among different genre, show us which genre have the most frequencies of rating:

- **Genre Distribution Across Ratings:** Observe the distribution of genres within each rating category. This helps understand the prevalence of genres in different rating ranges.
- **Dominant Genres:** Drama has the most recorded ratings.
- **Genre Preferences for High Ratings:** Genre like Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Sci-fi, KidsFamily tend to have consistent high ratings which takes part half of each total ratings.
- **Comparisons Across Ratings:** For well distributed ratings, that genre tend to have fastidious viewer. For instance, Horror and Thriller have the same distribution among ratings which indicates they only sastified small point viewer.

Figure 7. Scatter plot of Ratings and Movie Duration

A scatter plot of rating and movie duration can provide several insights into the relationship between these two variables:

- **Overall Trend:** Most film are distributed in duration from 75 minutes to 175 minutes.
- **Correlations:** The Duration has small impact on what score the Rating is.
- **Viewer Preferences:** Understand viewer preferences by observing where the majority of highly-rated movies fall on the duration spectrum. This can provide insights into audience expectations and preferences.

A bar chart depicting the highest ratings for directors within a particular genre can provide valuable insights into which directors excel in specific genres:

- **Top Directors by Genre:** Identify directors who consistently receive the highest ratings within a specific genre. This information can be valuable for understanding the

association between directorial expertise and audience appreciation for particular genres.

- **Audience Preferences:** Gain insights into audience preferences by observing which directors contribute to the highest-rated movies in each genre. This information can help studios and filmmakers understand audience expectations and tastes.
- **Directorial Impact on Ratings:**Assess the impact of directors on movie ratings within different genres.
- **Variation Across Genres:** Explore how directors' highest ratings vary across different genres. Considering the ability of directors impact on related genre.

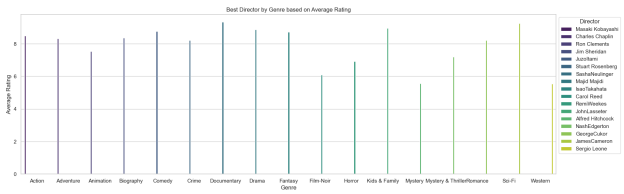

Figure 8. Bar chart for Highest rating for Director in particular Genre

IV. Model selection and architecture

While investigating the problem, we dived into two main ideas. Firstly and traditionally, we use some common machine learning models such as BERT, LSTM, etc. This gives us an easier approach, but the results will be predictable. Hence, we decide to use the second approach, along with these recent development of AI, sentiment analysis using Large Language Models (LLMs).

While choosing a model for this task, we searched in Open LLM Leaderboards [1] and find that, LLaMA-2 7B is the best suitable for our project since:

- LLaMA-2 is on the top of best accuracies model recently. Its tokenizer and pre-trained knowledge are nearly completed. This helps us save time in continual pre-training for this model to read "normal text" better.
- LLaMA-2 has an incredibly high MMLU (Massive Multitask Language Understanding [2]) score and it will have a good initial performance on understanding our requirements.
- LLaMA-2 gives us a 7 billions parameters version. Despite the small number of parameters, this versions has an acceptable accuracy along with shorter time consumed for fine-tuning.

For a clear understandings about the principles of LLaMA-2 and how it works, we will discuss a small section below with citation from the authors of the model, MetaAI [3], [4].

LLaMA is a new model generation developed by Meta AI, using Transformer algorithm. They train large transformers on a large quantity of textual data using a standard optimizer.

However, Meta AI has altered the original architecture with some state-of-the-art techniques, and here are some noticable changes:

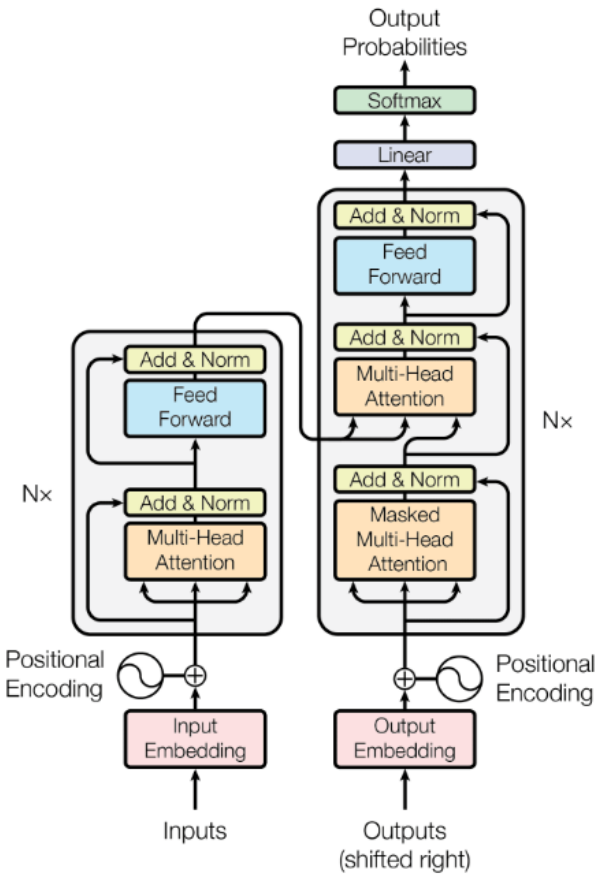

Figure 9. The encoder-decoder structure of the Transformer architecture, taken from "Attention Is All You Need"

- **Pre-normalization [GPT3].** To improve the training stability, they use the RMSNorm normalizing function to normalize the input of each transformer sub-layer, instead of normalizing the output.
- **SwiGLU activation function [PaLM].** They replace the ReLU non-linearity by the SwiGLU activation function, to improve the performance. However, they change the dimension from $4d$ in PaLM to $\frac{2}{3}4d$.
- **Rotary Embeddings [GPTNeo].** They remove the absolute positional embeddings, and instead, add rotary positional embeddings (RoPE) at each layer of the network.

V. Model fine-tuning

In model fine-tuning, we decide to use LoRA fine-tuning instead of full fine-tuning because of the small dataset. With the dataset we have, the use of full fine-tuning will lead to catastrophic forgetting. Also, with the given dataset, it is better for us to use supervised fine-tuning using SFTTrainer module get a higher performance.

Suppose we are given a pre-trained autoregressive language model $P_\Phi(y|x)$ parametrized by Φ. Consider adapting this pre-

trained model to downstream conditional text generation tasks. Each downstream task is represented by a training dataset of context-target pairs: $\mathcal{Z} = \{(x_i, y_i)\}_{i=1..N}$, where both x_i and y_i are sequences of tokens. During full fine-tuning, the model is initialized to pre-trained weights Φ_0 and updated to $\Phi_0 + \Delta\Phi$ by repeatedly following the gradient to maximize the conditional language modeling objective:

$$\max_{\Phi} \sum_{(x,y)\in\mathcal{Z}} \sum_{t=1}^{|y|} \log(P_\Phi(y_t|x, y_{<t})$$

One of the main drawbacks for full fine-tuning is that for each downstream task, we learn a *different set* of parameters $\Delta\Phi$ whose dimension $|\Delta\Phi|$ equals Φ_0. This may lead to catastrophic forgetting.

However, in LoRA techniques, we hypothesize the updates to the weights also have a low "intrinsic rank" during adaptation. For a pre-trained weight matrix $W_0 = \mathrm{R}^{d\times k}$, we constrain its update by representing the latter with a low-rank decomposition $W_0 + \Delta W = W_0 + BA$ where $B \in \mathrm{R}^{d\times r}, A \in \mathrm{R}r \times k$ and the rank $r \ll \min(d, k)$. During training, W_0 is frozen and does not receive gradient updates, while A and B contain trainable parameters.

When applying LoRA to Transformer architecture, we have for weight matrices in the self-attention module $(W_q, W_k, W_v, W_o))$ and two in the MLP module. We treat W_q as the single matrix of dimension $d_{\text{model}} \times d_{\text{model}}$, even though the output dimension is usually sliced into attention heads. We limit our study to only adapting the attention weights for downstream tasks and freeze the MLP modules. [5]

Along with LoRA fine-tuning, we considered and chose to apply quantization-aware training techniques to make our models can run directly on some free platform such as Google Co-laboratory (1 x T4 GPU). This called QLoRA techniques. In QLoRA techniques, we apply the block-wise k-bit quantization formula: [6]

$$X^{\text{int8}} = \text{round}\left(\frac{127}{\text{absmax}\left(X^{\text{FP32}}\right)} X^{\text{FP32}}\right)$$
$$= \text{round}\left(c^{\text{FP32}} \cdot X^{\text{FP32}}\right)$$

where c is the quantization constant or quantization scale.

VI. Implementation

Here, with the original model, LLaMA-2 cannot understand clearly and do the sentiment analysis task properly and do not have many application in real life. So, we decide to fine-tune the model twice, which are:

- Select only film review and its sentiment (Positive or Negative). This helps LLM to understand the attitude of the input. After this stage, LLM can understand the attitude and give the score for some classfied category of a film review.
- Give LLM the film name and all the collected review contents. By doing that, LLM will have a knowledge about film ratings and users can ask LLM to give the ratings of the film by typing its name, rather than pasting each reviews for LLM to read. *(curently development)*

Our model is fine-tuned with 4 x V100 machine. It took about 1 hours for each training cycle. All the codes for our projects are uploaded into out Github repository and our model is published on our HuggingFace repository here.

VII. Summaries and Improvements

After running several times, we get the accuracy of the chosen datasets is represented in the table below. [1]

Model name	Accuracy	MMLU	TruthfulQA
LLaMA-2 7B Original	0.45	46.87	38.76
Falcon 7B Original	0.36	27.79	34.26
LLaMA-2 7B Fine-tuned	0.65	48.32	38.51

Here, we can see that, with the movie review sentiment analysis task, our model outperform the others, that proves our methods have a positive effects on models. From that, we aims to give an assistant for producers to investigate more about audience's attitude, also, help audience to get the film review for the next time they go to the cinema.

Acknowledgments

This work is supported and supervised by Professor Khoat Q. THAN, Trung V. TRAN, Oanh T. NGUYEN, and Mai-Anh BUI under the course Introduction to Data Science. Also, our research and project are partially supported by other professors and students at the School of Information and Communication Technology, Hanoi University of Science and Technology.

References

[1] Beeching, E., Fourrier, C., Habib, N., Han, S., Lambert, N., Rajani, N., … Wolf, T. (2023). Open LLM Leaderboard. Retrieved from https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

[2] Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., ... & Han, J. (2023). Don't Make Your LLM an Evaluation Benchmark Cheater. arXiv preprint arXiv:2311.01964.

[3] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[4] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[5] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

[6] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314.