DatA science, Learning and ApplicationS
UM4IN814

Project
Final Report

Manga/light novel/anime, the unofficially translation problem in Vietnam, challenges and chances.

(*Tran Quoc An TRUONG – 21402950 – BIM* - Github)

1. Abstract:

Manga (Japanese comic), lightnovel (Japanese roman) or the other Japanese media products are higly popular in Vietnam from 00s especially in the decades of internet exploding. Many websites had translated and published uncountable amount of manga from then. Including me, my friends, and others GenZ Vietnameses, I could say about 90% of us had read at least one manga. Litterally, we had grownth up with them. But, there is problem about the translation or publicity copyright. While many book store and others distributors have bought the license many manga/lightnovel/anime but the amount of people who consume these products which have been translated unofficially is still numerous.

In this project, we will investigate and response to questions:
- Why Vietnamese choose unofficial or illegal manga/lightnovel/anime translation product over the official ones?
- How about the current revenue of the official print partners? Is it good?
- Is there any way to tackle it? Is this a chance that we can take advantage to increase the revenue?

How:

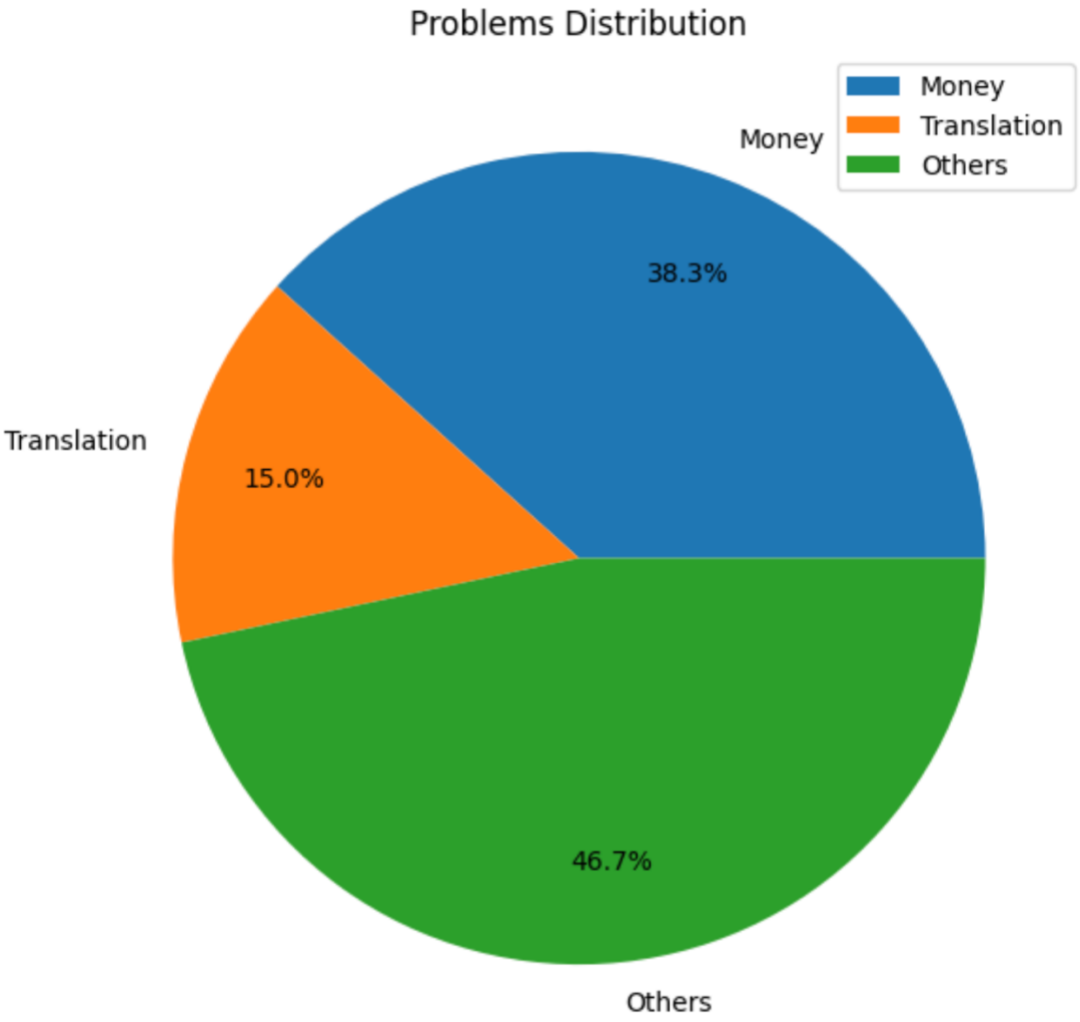
- Data collecting:
  - Crawl data from Facebook comments to see why they choose to read manga/ln(s) unofficially translated over official ones. (Almost in Vietnamese): Using Facebook Posts Scrapper
  - Crawl selling data of Bookstores from the online marketplaces (Shopee, Lazada, Tiki, etc.): A little tricky because of Shopee required authenticated to see their websites. So, I will take screenshot manually then pass to an VLM (Language-Vision Model) to extract the needed information (book names, prices, selling count).
  - Crawling reading/watching count from unofficial manga/light novel/anime unofficial publisher.
  - GDP Vietnam and Vietnamese purchase's power
  -
- EDA:
  - Use a LLMs to translate comments to English.
  - Refine the selling data (datatype, missing data, book name, etc.)
  - Classify comments (using LLM to text segmentation by given classes i.e. reasons) and visualize the distribution to see which is the most frequent reason of consuming unofficial product.
  - Visualizing the distribution of consuming official/unofficial products.
  - Vietnamese people's goods buying distribution (food, cloths, book, etc.) and official published Manga/light novel/anime consume distribution to analyze Vietnamese Japanese cultural product's purchase power.
  -
- Prediction model – Linear Regression:
  - The average monthly income of urban area in order to find the optimal price for these cultural products especially manga in Viet Nam.
- Propositions for the problem:
  - Apply the model like a marketplace where any translation group can distribute their work directly on the official reading website of official publisher.

- Therefore, customer can freely choose their loved translation group to read and increase the variation of manga/light novel/anime for customer to choose.

2. Data Presentation & Description

2.1 Facebook comments:

- **Text Data** extracted from public Facebook groups dedicated to Anime/Manga/Light Novel consumption in Vietnam because the most pratical in Vietnam is Facebook, about 72mil active user for 101mil of population. This provides **direct qualitative evidence** of consumer sentiment and rationale.
- However, it's high risk of **Sentiment Bias** (comments tend to be emotionally charged, negative, or polarized) and limited to users active on specific Facebook groups.
- Due to the high security and anti-scraping measures of Facebook, data is collected manually: copying comment threads into local text files then using hard code to clean and filter out only the comments.
- Then, apply a LLM (Ollama in this case) to translate the original text from Vietnamese to English for analyzing and store all original and translated comments into a Pandas DataFrame.
- Lastly, by using builtin string supported function on DataFrame column to count the frequency of given word in the translated comment, those given words are related to the 3 principal problem classes (for example: money, poor, condition for *money problem*, translation, translate for *translation problem*, and the rest for *Others* like the delay from publisher, lacking the licenced publisher for desired titles, etc.)

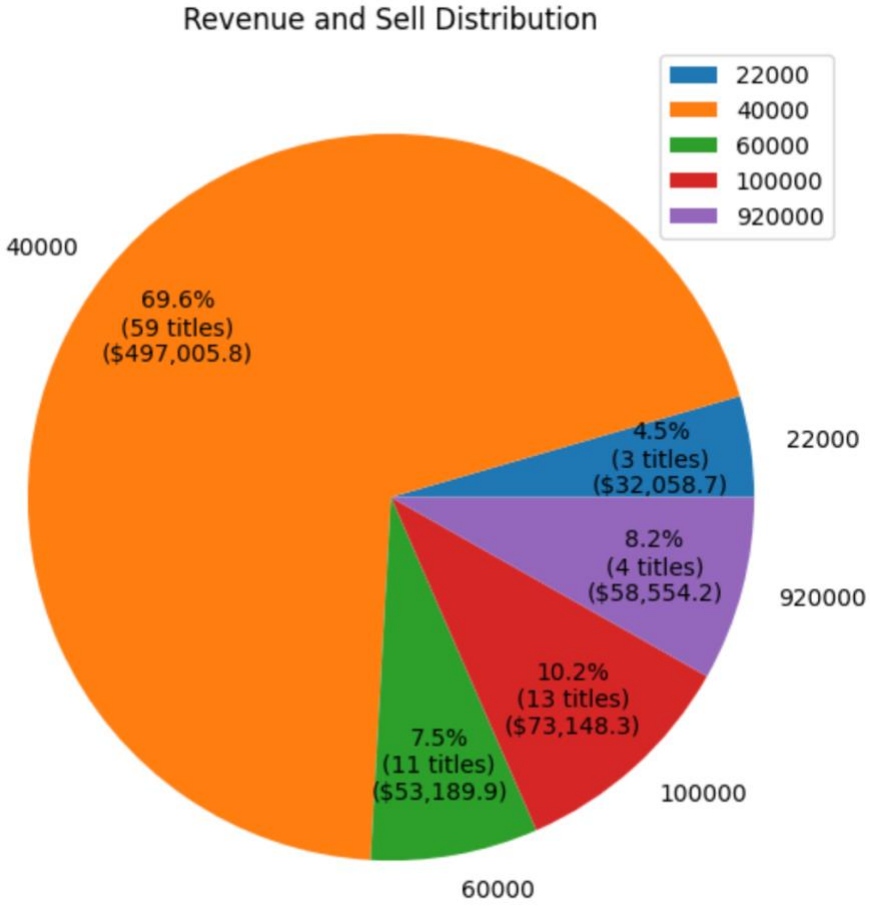


- As we can see, aside Others problem about publisher dont accquire the rights or delay in publishing, *money* is definitively unavoidable problem here which appeared in 46,7% of all of comments while in fact, the translation is a bit less important.

2.2 Selling Data of Official Kim Đồng Publisher from Shopee marketplace:

- Data collected from the official store of **Kim Đồng Publishing House** on the Shopee marketplace. Shopee's dominant position in Southeast Asian e-commerce validates it as the primary channel for legal purchase transactions in Vietnam. It also justify directly the attractiveness of Japan media and the consuming power of Vietnamese Manga lovers in legal market, serving as the counterpoint to the unofficial view counts.
- This data provides also the official pricing data and the observed outcome for the price prediction regression model.
- Due to the strong anti-bot and anti-crawling defenses on Shopee, direct scraping is not feasible. The workaround involves **manually taking screenshots** of product listings and then processing these images using a **Vision Large Model (VLM)**. Discretly, I used the free tier API from Gemini CLI to OCR the images captured manually on Shopee.
- **Cleaning Required:** The raw output from the Gemini CLI model is good enough, thus I just have to ensure the datatype of numeric features is integer. Then I proceed to data exploration and analysis by finding the *max price* is 920.000 VND while the *min price* is 22.000 VND. Then by grouping those titles by price ranges (from 0-20.000, 20.000-

40.000, 40.00-60.00, 60.000-100.000 and 100.000-920.000), I visualize the Revenue and Selling Distribution as below:

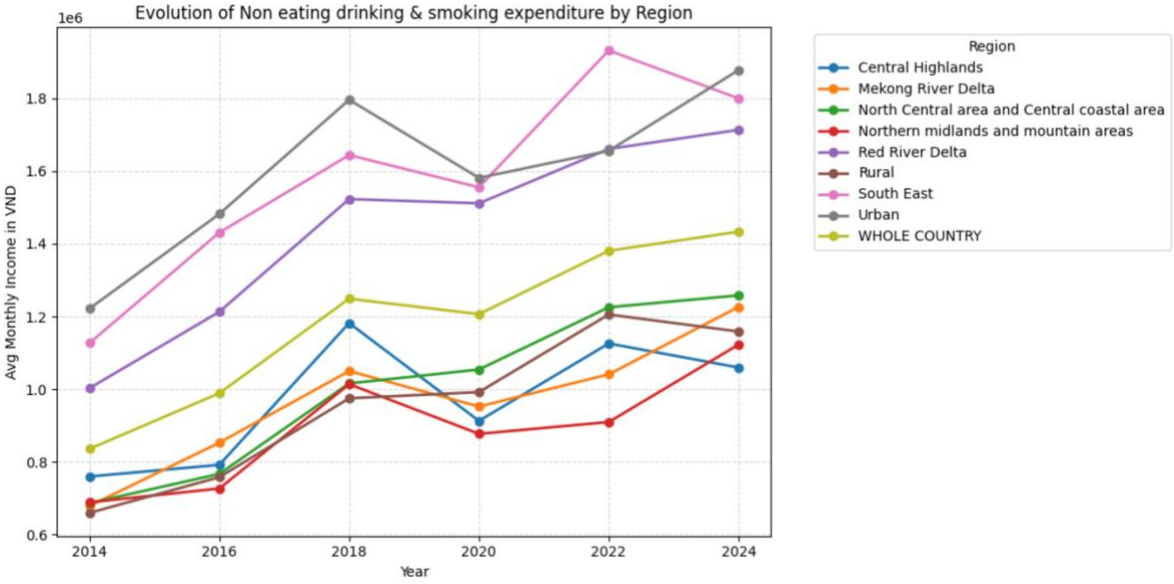


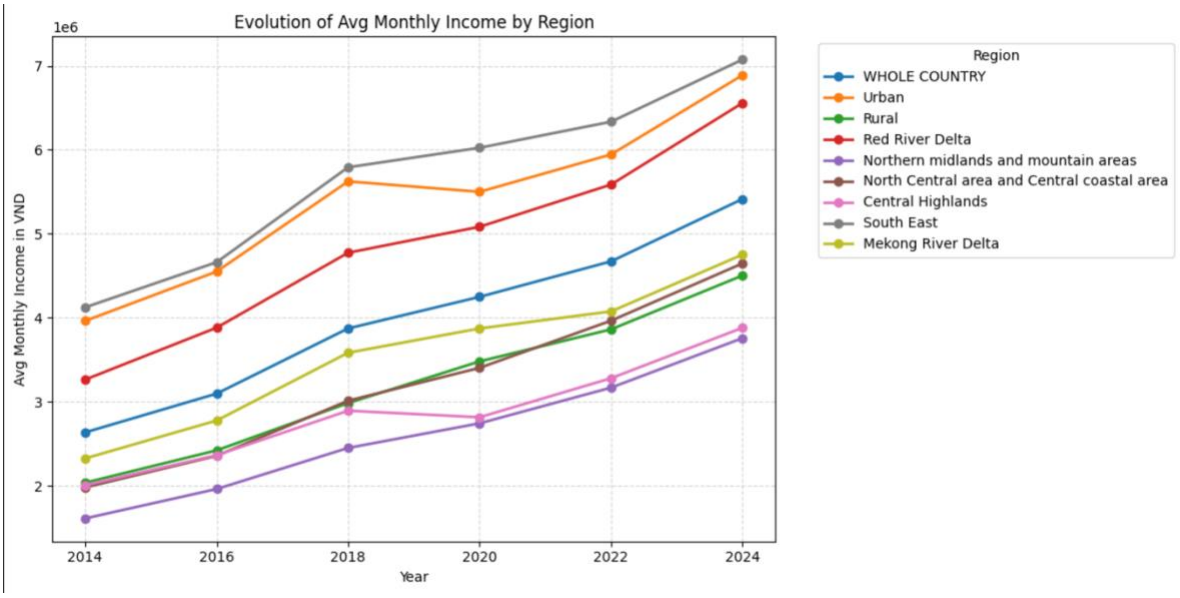
- In the graph, numbers on the legend are the prices bins for collected titles of Kim Đồng on Shopee. As it shows, majorities of those titles are placed in range from 29.999 to 40.000 make the most profit ($497.000 ~ **69.6%)** while ones less than 20.000 and higher than 100.000 are least profitable.
- As we can see, only 90 titles collected but that make half a million dollars of profit for licenced titles, because there are much more titles that are currently attracting the young people.

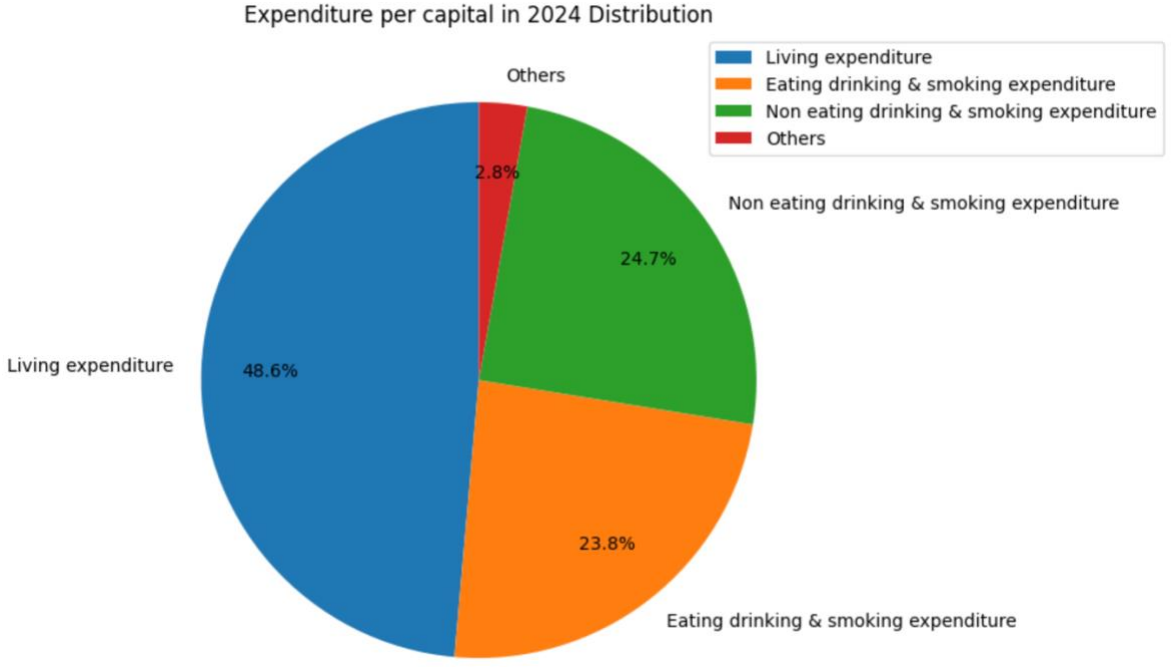
2.3- National Statistical Office of Vietnam Dataset:

- This data is normalized and provides an insight about economic context of Vietnam and most important the purchase goods power of Vietnamese people. This dataset contains the distribution of expenditure per capita monthly by type of goods, the average income per capita from 2014 to 2024.
- It represents the capability the growth of revenue of Vietnam market.
- The extensive time-series data (2014-2024) will be used to **predict future incomee**.
- **This** data is organized as a multi-index pandas Data Frame with:
  - **Column structure**: Two-level Multi-Index:
    - Level 1: Years (2014, 2016, 2018, 2020, 2022, 2024) - 6 time points
    - Level 2: Variables (5 types of expenditure, average monthly income) - 6 features per year
    - Total columns: 36 (6 years × 6 variables)
  - **Row structure**: 9 rows representing geographical regions within the nation

    - **Data dimensions**: 9 × 36 Data Frame
- The data distributed by goverment so the quality is mostly clean, I just have to merge the Expenditure DataFrame to Average Monthly Income DataFrame, and remove a NaNs row caused from loading the original csv files.
- For Prediction model:
    - **Model Selection:**
        - To predict both *average monthly income* and *noneating expenditures*, I'm planning to train 2 Multivariable Polynomial Regression **separately**, because they can learn the non-linear trend in economic data, moreover, there are 4 4 random variables and 2 dependent variables for each year so that the multivariable is reasonable.
    - The prediction targets are:
        - **Monthly income** for 2026 (next year)
        - **Monthly expenditure** for 2026 (next year)
        - **Format**: Continuous numerical values (2 outputs)
        - **Dimensions**: 9 × 2 (predictions for 9 regions)
    - Training Data and Validation Strategy:
        - **Volume**: With only 1 geographical region, 6 time points and 4 features of each year, the dataset is truly limited (24 total feature-year observations).
        - **Time-series**: Train on years 2014-2020 and test on data of 2022 and 2024.
        - This respects temporal ordering and tests the model's ability to forecast the most recent year.


Evolution of Non eating drinking & smoking expenditure by Region
1e6
Avg Monthly Income in VND
Year
Region
Central Highlands
Mekong River Delta
North Central area and Central coastal area
Northern midlands and mountain areas
Red River Delta
Rural
South East
Urban
WHOLE COUNTRY


Evolution of Avg Monthly Income by Region
1e6
Avg Monthly Income in VND
Year
Region
WHOLE COUNTRY
Urban
Rural
Red River Delta
Northern midlands and mountain areas
North Central area and Central coastal area
Central Highlands
South East
Mekong River Delta


Expenditure per capital in 2024 Distribution
Living expenditure
Eating drinking & smoking expenditure
Non eating drinking & smoking expenditure
Others
2.8%
24.7%
48.6%
23.8%

- Through these figures shown above, we can see the Vietnamese economic rapid growth in the period of 10 years. Besides, the Vietnamese people tends to invest more and more their income into non eating expenditure like book and multimedia (**24.7%**) especially in the Urban and South East regions where they spend over 1.800.000VNĐ and they are also the regions that have the highest average monthly income nationwide. Therefore, I concentrated on building the average monthly income and non eating expenditure models for Urban region.
- By using just a simple Rigde Regression Linear model on very limited dataset (2014, 2016, 2018, 2020 for training and 2022, 2024 for testing with features are all the expenditure types except Non-eating one) the result is as expectation that does not perform very well. Using, RMSE, the predictions from 2 models shows their poor performance (45,75 and 7.27 respectively).

2.4 Unofficial Anime Streaming Website Anime List:

- Data collected directly from a popular Vietnamese anime streaming website.
- This data gives an insight about which animes are on trending.
- This website is highly protected from crawling data bot or framework by **Cloudflare Captcha**. So, the top 40 animated film series most trending of the year 2025 and their views were recorded manually.
- This data is essential for the **Content Gap Analysis**: it directly informs which titles the official market (Netflix/Kim Đồng) is failing to acquire quickly enough to satisfy local demand.
- The watching count from Unofficial Anime Streaming website data is not time framed. Therefore, some of them are released from a long time et they get very high views (over 20 million) while the total population of Vietnam is only about 101 mil.
- Besides of this website, there exist several website for streaming pirated contents (including anime or Japanese animated also).

2.5 Official on-aired Anime on Netflix:

- This data which is open source and reliable on Kaggle, contains current available TV shows and movies on Netflix, updated every month.
- **Key Limitation:** Most public Netflix datasets are compiled by third parties (like Flixable or JustWatch) and are often **globally scoped** or specific to the US region. They are **not guaranteed to be accurate for Vietnamese regional licensing**.
- Serves as the critical benchmark to determine **Content Gap**. By comparing this list to your Scraped Demand (Unofficial Views), you quantify *which* high-demand titles Netflix has **failed to acquire** for the Vietnamese market.
- Downgrade of this data is limited **geographically**. In fact, the on-aired shows avaibility is slightly changed by location. That said, in this case, we study only in Vietnam while the dataset seems USA/EU centric, that might affecte the result.
- In order to highlight the *Translation* and *Publisher* -related problems that are indicated by the Facebook comments, I compared the current on-trend anime from unofficial streaming platforms, to licenced one, the result is quite transparent to the 2 mentioned problems.

```
(['SAINT SEIYA: Knights of the Zodiac',
  'A Silent Voice',
  'A Whisker Away',
  'A.I.C.O.',
  'Aggretsuko',
  'AJIN: Demi-Human',
  'Akame ga Kill!',
  'Angel Beats!',
  'Anohana: The Flower We Saw That Day',
  '忍者ハットリくん'],
 173)
['One Piece',
 'Black Clover',
 'Detective Conan',
 'Jujutsu Kaisen 2nd SS',
 'Bleach',
 'Demon Slayer - Swordsmith village Arc',
 'Demon Slayer - Hashira Training Arc',
 'The Eminence in Shadow',
 'Soul Land - Douluo Dalu',
 'Tsuchimiki: Mônlit Fantasy']

{'Bleach', 'Fairy Tail'}
```

- The first list is belonged to Netflix, the second one belonged to the pirated streaming platform and the last one is the result of comparison of these 2 lists. As we can see, there are only 2 animated contents that are on-trend on pirated platform appear on the Netflix, despite of the amount of anime on Netflix is vast but it does not satisfy demands from potential users. Thus, this has driven those potential users to find their loved, desired animated on the unlicenced platforms.

3. Ethical & legal aspects in Data Collection

- There are some considerations about the legality of my dataset, especially the selling data from Shopee, unofficial anime streaming website and comments from Facebook public posts. These data were obtained via unconventional methods, necessitating a formal risk assessment against privacy and platform policies i.e. *Terms of Service* (**TOS**) and *Anti-bot Policy*.
- Since this project is **academic**, **non-commercial**, and by focusing exclusively on **metadata** (titles, prices, view counts, comments) and **avoiding the collection of the actual copyrighted content** (manga pages, video streams), the project minimizes the risk of direct copyright infringement.
- The project presents minimal ethical risk regarding personal privacy because of the nature of the collected data. No *Personally Identifiable Information* (**PII**) of individual

consumers (such as IPs, device IDs, or private login credentials) is gathered. The project is compliant with general data protection principles (like those behind **GDPR**).

- Therefore, the data acquisition methods carry **TOS risk** but asserts that the data is **ethically safe** for use.

4. Conclusions

- Through this project, we have identify the main problem (*Money*) from the possible causes from **Extracted Facebook comments** dataset. However, the *Translation* and *Others* problem can not to be casted aside, especially when comes to those streaming pirated animated contents platforms.
- Thank to the **NSO Dataset**, we can observe the rapidiness in economy growth of Vietnam based on the increasing of *average monthly income*, and also the percentage of *non-eating expenditure per capita* over regions especially South-East and Urban.
- Besides, there is the **failure** in data preparation for prediction models that is the amount of data is too few, and the features does not seem too much related to the average income prediction.Thú, that makes the model is likely unusable.
- On the other hand, with this growth rate of Vietnamese economic, it proves that Vietnamese people has capability for buying the licenced products, it's just because the lack of diversity of those products.
- Furthermore, with this analysis result, I think the money problem will gradually resolved but the T*ranslation and Others* are needed to be get much considerations from the original publisher in Japan. In fact, I'm also an Japanese culturals contents lover, I do really understand this Paradox that many customers are willing to byt the copyrights products from licenced publisher/vendors but they cannot find them anywhere except those pirated content publish platform. There are also some credits for Muse VN, AniOne VN for have purchasing many copyright of animateds and streaming them freely on Youtube in exchange for Ads but those animated are not enough and plenty of user have high demand for the other animated especially classic or vintages ones which are difficult to accquire theirs licence.
- This problem seems to be happening in France for now, where the diversity of anime and manga is the main problem for pirating contents.
- Despite of the effort of developping the manga reading online app from the largest Japanese publisher Shueisha, the *Translation* is still present, not only for Vietnamese, French but also other country than English native speaker ones.
- Therefore, I suggest an online platform that open to third party translation teams, which are the core of those pirated platforms. They are doing with their love but only some donates in exchange. Those Japanese can give them the right to translate their products, the reader gives the reviews and the translation will get compensated based on the views of theirs translations. It's a arrow for 2 bỉrds, that will not only reduced the publishing pirated contents but also get more populare and expand the market share globally.