**Automatic Summarisation of TripAdvisor**

**Reviews of Spanish (Madrid) Restaurants**

**Graduation Thesis: The Interim Report**

Dan Luo

Student ID:50079757

Program: CS

[d.luo.21@aberdeen.ac.uk](mailto:a.ciocarlan.12@aberdeen.ac.uk)

*Aberdeen-SCNU Joint Institute of Data Science and AI,*

*University of Aberdeen, Aberdeen AB24 3UE, UK*

**Introduction**

The project, **"Automatic Summarisation of TripAdvisor Reviews of Spanish (Madrid) Restaurants"**, addresses the need for concise and efficient ways to summarize user reviews on TripAdvisor. Given the vast amount of user-generated content on platforms like TripAdvisor, manual browsing through reviews becomes cumbersome for users seeking quick and relevant insights. Automatic text summarization, powered by natural language processing (NLP) techniques, provides a scalable and user-friendly solution.

This project explores both the challenges and opportunities of generating automatic summaries using state-of-the-art models like BART (Bidirectional and Auto-Regressive Transformers). By leveraging real-world datasets and cutting-edge technologies, this project contributes to improving user experience in online review platforms.

**Goals**

The primary goals of this project are as follows:

* Develop a text summarization tool that can automatically generate concise and meaningful summaries for restaurant reviews.
* Use advanced generative NLP models, such as BART, to ensure high-quality summaries.
* Deploy the tool as a web application for easy access and usability.

Secondary or optional goals include:

* Incorporating sentiment analysis into summaries to provide users with a clearer understanding of review sentiments.
* Enhancing the summarization model with multilingual support, particularly for Spanish and English.

**Methodology**

To achieve the project objectives, the following steps will be taken:

* **Literature Review**:
  + Study existing work on extractive and abstractive text summarization techniques.
  + Explore how platforms like TripAdvisor manage and present user-generated reviews.
* **Dataset Preparation**:
  + Clean and preprocess the TripAdvisor dataset containing reviews of Madrid restaurants.
  + Split the dataset into training and testing subsets for model evaluation.
* **Model Implementation**:
  + Use pre-trained BART or T5 models and fine-tune them on the prepared dataset.
  + Experiment with hyperparameter tuning to optimize the summarization results.
* **Prototype Development**:
  + Build a Streamlit-based web application to demonstrate summarization capabilities.
  + Allow users to upload reviews or input text to generate summaries dynamically.
* **Testing and Evaluation**:
  + Evaluate the model performance using metrics like ROUGE and BLEU.
  + Gather user feedback to assess the practicality of the generated summaries.

**Resources Required**

The resources essential to the successful delivery of the project include:

· **Hardware**:

* A PC or laptop with at least 16GB RAM and a GPU (preferably with CUDA support) for faster model training and inference.

· **Software**:

* Python 3.8 or higher.
* Libraries: Hugging Face Transformers, Pandas, Streamlit, NLTK, and PyTorch.
* A code editor (e.g., VS Code, Jupyter Notebook) and a web browser for application deployment.

**Risk Assessment**

· **Potential Risks**:

* The pre-trained models might require more computational power than available, slowing down training or inference.
* Dataset issues such as imbalanced reviews or noisy data could affect the summarization quality.
* Deployment challenges in creating a responsive and efficient web application.

· **Mitigation Strategies**:

* Use cloud-based platforms like Google Colab or AWS for computationally intensive tasks.
* Perform extensive data cleaning and augmentation to ensure a robust dataset.
* Optimize the web application by minimizing model size or using an API-based approach.

**Timetable**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Literature review |  |  |  | |  | |  | |
| Dataset preparation |  |  | | |  | |  | |
| Model implementation |  |  | | |  | |  | |
| Web Application Development |  |  | |  |  | |  | |
| Testing and evaluation |  |  | | |  |  |  |  |
| Final Report Writing |  |  | | |  | |  | |
|  | January | February | | | March | | April | |

**References**

[1] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.