Team Champagne project documentation

Sentiment Analysis of Google Reviews

ZongYun Li(netID: zyl2), Chi Han(netID: chihan3)

December 7, 2022

1) An overview of the function of the code

a)project.ipynb

- Data collect: Use request to collect data from 6 restaurants' google comment.
- Data append & balance: Append the data from 6 restaurants, and let negative data and positive data become same amount.
- Model training: Use balanced data to train and test bert model, and use labeled data to train and test reberta model.

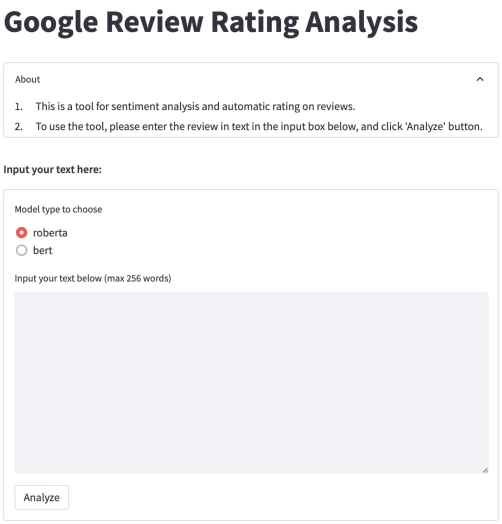
b) Demo Interface (Figure 1)

- Sentiment Analysis: our tool provides a predicted label of the text being either "positive" or "negative", as well as the estimated distribution of labels.
- Similar Review Retrieval: our tool also retrieves similar review texts from the training corpus and list them in the results. This provides users with a better understanding of the rationales behind the prediction.

2) Software implementation

a)implementation

In this project, we first used request to collect the data from google map's comment. The data size is 10220. We also labeled the data manually because there may be some wrong rates. We used the labeled data in bert and reberta model, but it didn't work well in bert model. To improve bert model, we make the negative data and positive data be the same amount. The accuracy of bert and reberta are 90.7% and 92.9%.



(a) Application page interface

(b) Format of results

Figure 1: An overview of the application.

For the Demo interface, we use Stremalit package[1] to build the webpage. In the backend, we use the trained model for label prediction results, and use soft-max operation to get the label distribution estimate. Finally, we use the cosine-similarity of sentence embedding features to retrieve most similar review texts.

b) future improvement

I think the data is insufficient. Collecting more data from the net will improve the model accuracy. By adding the reviewers' gender feature, it may help model predict more accurate. This is because different gender have different kind of expression.

3) How to install and run a software

Prerequisites

The following python packages are required for running our tool:

- streamlit (https://streamlit.io/)
- pytorch (https://pytorch.org/)
- transformers (https://huggingface.co/docs/transformers/index)
- simpletransformers (https://simpletransformers.ai/)
- scipy

[1] `https://streamlit.io`

- pandas

Installation

First, git clone our repository:

```
git clone https://github.com/same8891/CourseProject.git
```

Then, go to the ‘code‘ folder:

```
cd CourseProject/code
```

Then, download these two Google Drive folders (our model parameters) to the current position, and rename them as `roberta` and `bert`, respectively.

- `https://drive.google.com/drive/folders/1ESOczJViLGB7yyctyFch6C-3EhcOoq-J`
- `https://drive.google.com/drive/folders/10Qj5SHwn_jMGDFWzRs3nDx9sTtNZF3p9`

Finally, launch the streamlit server:

```
streamlit run app.py
```

You should expect the following outputs in the terminal:

```
You can now view your Streamlit app in your browser.
Local URL: http://localhost:8501
```

Network URL: `http://10.186.103.205:8501`

If the webpage is not automatically popping-up, please copy-paste the url to your browser and open.

4) Group division

ZongYun Li: Collect the data from google comment; Labeled the data; Train bert model and reberta model by simpletransformers.

Chi Han: labeled the data; implemented the demo application codes for sentiment analysis and similar review retrieval; prepared the presentation slides and video.