Spoiler Alert: Using Deep Learning to Detect Spoilers in Text and Generate Similar Spoiler-Free Text

Abolfazl Eshagh, Mobina Salimipanah, Parnian Razavi, Ali Nikkhah, Sarina Zahedi, Ramtin Khoshnevis

Abstract

……
……
………………………………………………………

1. Introduction

……………………………………………………………………
………………………………………………
……………………………………………………………………
…………………..
………………………………………………..

2. Related Work

……………………………………………………………………
……………………………………………………………………
…………………………………………………………….

3. Datasets

……………………………………………………………………
……………………………………………………………………
……………………………………………………………………
……………………………………………………………………
……………

4. Method

Dataset preprocessing

……………………………………………………………………
……………………………………………………………………
………………………………………………………………..

4.1. Classification

……………………………………………………………………
…………………………………………………………

4.2. Generation

In the generation phase, we tackle multiple tasks, primarily focusing on generating spoiler-free plot summaries from the original plot, which contains spoilers, in our main dataset [reference needed]. Additionally, we generate spoiler-free review summaries from the initial review texts in the dataset that contain spoilers.

Spoiler-free plot summary generation
First, we performed initial preprocessing on our dataset [reference needed]. It is important to note that some entries lacked values in the '`plot_synopsis`' column, necessitating the removal of those samples. We then defined our model using the '`BartForConditionalGeneration`' [reference needed]. This model was fine-tuned over 10 epochs. Unfortunately, due to the limited size of our available dataset, the model quickly overfitted, making further fine-tuning beyond 10 epochs impractical. The best state of the model, captured during this training session, was saved and utilized for evaluation. The results for this task are evaluated using the metrics outlined in Section 5 and are presented in Table 2.

Spoiler-free review summary generation
Similar to the plot summary generation process, we began by preprocessing this part of the dataset [reference needed]. The only two features needed were the '`review_text`' and '`review_summary`' columns, so all other columns were discarded. It is worth mentioning that this part of the dataset is significantly larger, containing more

1

than 450,000 samples of review text and corresponding summaries. Due to computational resource constraints, we had to downsample this dataset randomly. Next, we loaded the '`BartForConditionalGeneration`' [reference needed] model and fine-tuned it on the prepared data. Unlike the plot summary task, the larger dataset size allowed the Bart model to continue achieving lower losses with each epoch. The results for this task are evaluated using the metrics outlined in Section 5 and are shown in Table 3.

5. Metrics

……………………………………………………………
……………………………………………………………
…………………………

6. Results

……………………………………………………………
……………………………………………………………
…………………………………

7. Future Works

……………………………………………………………
……………………………………………………………
……………………………………………………………
……………………………………………………………
…………

8. Conclusion