Sentiment Analysis on Steam Game Review using RoBERTa: Does Sarcasm Matter?

Yi Hung Chen(yihch883)
732A81 Text Mining
yihch883@student.liu.se

Abstract

Sentiment analysis on customer reviews is a widely used Natural Language Processing(NLP) application. However, some types of reviews, such as video game reviews, might contain sarcasm, which makes the traditional NLP model less than ideal. This paper aims to explore the state-of-the-art RoBERTa model's performance on the "Steam Review Dataset", which Steam is one of the biggest online video distributor. Also, four Fine-Tuned versions of the RoBERTa model using this specific dataset are tested. The performance of the Fine-Tuned model does improve over the base RoBERTa model and gives a more balanced result. On the other hand, the manual analysis of this paper suggests that although Fine-Tuned Roberta models achieve higher performance, they are not directly associated with the improvement to counter the effect of sarcasm.

1 Introduction

1.1 Motivation

Sentiment analysis has become a vital tool for getting insight into how consumers reflect on different products. This is particularly important in the digital entertainment industry. However, online reviews, such as game reviews on video game distributor "STEAM," can be hard to classify when using traditional Natural Language Processing (NLP) approaches such as Naive Bayes due to its dependency on context(Tan et al., 2021). Also, video game reviews can sometimes be sarcastic and ironic, which further decreases the performance of traditional classifiers (Xia et al., 2023) (Majumder et al., 2019). Hence, more complex solutions such as RoBERTa can be utilized.

1.2 Goals

This project will first look into the original RoBERTa model and check the accuracy on the Steam Review dataset. Then, the project will proceed to fine-tune a RoBERTa model based on the Steam Review dataset. Finally, the manual analysis will be applied to the result of the fine-tuned model to examine if the misclassified reviews are due to sarcasm.

2 Theory

2.1 RoBERTa Model

RoBERTa model, which stands for "Robustly optimized BERT approach", is an improved version of the well know NLP model BERT(Bidirectional Encoder Representations from Transformers)(Devlin et al., 2018).

RoBERTa was developed with the goal of improving efficiency and accuracy over BERT model. These improvements are done by using larger training data and bigger batches, removing the next sentence prediction objective from the original BERT model, dynamically changing the masking pattern, etc. This makes RoBERTa model gain significant accuracy improvements on benchmarks such as GLUE, RACE, and SQuAD(Liu et al., 2019).

2.2 Fine-Tuning RoBERTa Model

In order to gain better performance on a specific task. RoBERTa model can be fine-tuned. One of the fine-tuning approaches is to train on specific datasets. This can be seen as transfer learning that utilizes the big, pre-trained dataset and trailer into the specific task. This is the BERT model family's nature, allowing the model to expand for downstream tasks.(Devlin et al., 2018).

3 Data

3.1 Data Description

The data that is being used for this project is "Steam Reviews Dataset 2021". The raw dataset contains 21.7 million reviews of different languages and from over 300 different video games. The data consists columns such as,Steam app id, App name, Review id, Language of review, Review

, Recommended, etc. In which, 44% of the reviews are English. Also 87% of reviews are positive(recommended).(M., 2021)

3.2 Data Usage

For this project, "App name," "Language of review," "Review," and "Recommended" are utilized. The two columns, "Review" and "Recommended," are used for training and classification. In which the "Recommended" column is used as a label where "True" is recommended and "False" is not recommended. The column "Language of review" is used to filter out unwanted languages, as this project will only focus on English reviews. Finally, the "App name"(Game title) is kept for further manual analysis.

Also, due to the computational limitation, this project will only utilize 1% of the English reviews, comprising over 96000 reviews.

4 Method

4.1 Data Pre-processing

In order to get the data that is wanted for this project, the data will follow the below pre-processing steps.

Step 1: Keep columns **"App Name"**, **"Language of Review"**, **"Review Text"**, and **"Recommended"**.

Step 2: Keep only entries where the **"Language of Review"** is English.

Step 3: Remove any rows containing missing values (NA) in the **"Review Text"**.

Step 4: Sample 1% of the data due to computational limitations.

Step 5: Map the values in the **"Recommended"** column to integers: True to 1 and False to 0.

Step 6: Convert all reviews to lowercase.

Step 7: Split the data into training (90%) and validation (10%) sets.

It should be noted that, in this project, the sample seed is fixed to ensure consistent results.

4.2 Analysis using RoBERTa

The Baseline model for this project is "roberta-base", which is access from Huggingface FacebookAI/roberta-base. This is the foundation model for the whole RoBERTa model family, and it is pre-trained and ready to use.

As the Baseline, the validation data will be used to evaluate the roberta-base model's performance. For the sentiment analysis pipeline, the pre-trained RobertaTokenizer, "roberta-base", is used to make sure the input review satisfies the model requirement. Truncation is enabled to make the input length consistent across the dataset. Also, the tokenized input is converted into PyTorch tensors. The sentiment predictions were mapped, with 'LABEL_1' indicating a recommended review(1) and 'LABEL_0' indicating a negative review(0).

4.3 Fine-tuned RoBERTa Model

To fine-tune a RoBERTa model. A custom dataset class is created to handle the pre-process step for training. The tokenizer,"roberta-base", is the same as the baseline model. Also, within the class, the reviews are converted into string format, and Labels are retrieved. Truncation and Padding are also enabled to make the sequence length consistent across the dataset. Finally, the tokenized input is converted into PyTorch tensors.

PyTorch's Dataloader object is also utilized to handle batching. For this project the batchsize is set as 8 due to GPU's Vram limit. For training data, "shuffle" is also used to shuffle training data at the beginning of each epoch to avoid the risk of model learning from the order of input.

For the training loop, the AdamW optimizer with a learning rate of 2e-5 is employed. Within each epoch, input IDs and attention masks are forwarded for each batch. Subsequently, the loss based on the batch's prediction is calculated, and via backpropagation, the optimizer adjusts the weight based on it. This process will continue until all batches are used, and the new epoch will start again.

In this project, four fine-tuned versions of the model are presented. Each version corresponds to a different number of training epochs: 3, 4, 5, and an extended version trained for 10 epochs. This setup allows for an analysis of the effects of additional training on the model's performance and the potential for overfitting.

Below are the hardware specifications used for training the model:

1. **CPU:** AMD Ryzen 7 5800X
2. **GPU:** NVIDIA GeForce RTX 3080 Ti 12GB
3. **Memory:** 32GB RAM

4.4 Model Evaluation

In addition to the overall accuracy, this project will focus on precision, recall, and f1-score when evaluating model performance.
Precision (Positive Predictive Value):

$$\text{Precision} = \frac{TP}{TP + FP}$$

where TP is the number of true positives and FP is the number of false positives.
Recall (Sensitivity):

$$\text{Recall} = \frac{TP}{TP + FN}$$

where TP is the number of true positives and FN is the number of false negatives.
F1 Score:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics are selected because of the imbalance nature between the two labels in the validation set. Therefore, using accuracy alone might provide misleading results.(He and Garcia, 2009) These metrics are done by using the Python module sklearn.metrics.classification_report.(Pedregosa et al., 2011)(Buitinck et al., 2013)

4.5 Manual Analysis

In order to understand if sarcasm is, in fact, compromising the model's performance, manual analysis is done on the misclassified review produced by the Baseline model and the better-performed fine-tuned model. The author inspects all misclassified reviews and labels the review based on the criteria below. If the review is sarcastic or seems very opposite to the recommendation status it gave, the review will be labeled as sarcasm(1). On the other hand, if the review is normal or the content is unrecognizable, it will be labeled as 0. "App name" is also considered alongside the review itself. This is because the author knows about some game's backgrounds, and some sarcastic comments can be spotted that way.

5 Result

This section will present the observation during the training step first. Then, the performance of different models will be examined. Finally, the manual analysis will be conducted after the performance evaluation.

5.1 Training Result

The accuracy and time are presented below in Table1. The training accuracy increases steadily as the epochs increase and the training time increases linearly. One observation is that the model already has high overall accuracy at 96% on the training set with only three epochs of training.

Table 1: Training accuracy and time

Epochs	Time(min:sec)	Accuracy
3	103:52	0.9653567137972461
4	138:17	0.9711094168745956
5	172:43	0.9760304038443767
10	346:11	0.9843822197578782

5.2 Base RoBERTa

For the baseline RoBERTa model, the performance is present in Table2. With an overall accuracy of 89%. However, as the previous section suggests, this may be due to the imbalance of the two labels. This is proven by examining the precision/recall/f1-score. The model has high performance on label 1(positive recommendation), with 0.89/1.00/0.94, respectively, on all metrics. In contrast, on negative recommendation (label), all metrics result in 0, which means the baseline model can be improved to gain better performance on the negative reviews.

Table 2: Classification Report of the Baseline RoBERTa Model

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1031
1	0.89	1	0.94	8588
accuracy			0.89	9619
macro avg	0.45	0.50	0.47	9619
weighted avg	0.80	0.89	0.84	9619

5.3 Fine-Tuned RoBERTa

The classification report of four Fine-Tuned Models are presented in Table3,Table4,Table4,Table6.

All models achieve similar results for label 1(recommended), with the f1-score all being 0.97. This

is a slight improvement over the baseline model. In contrast, all fine-tuned versions have higher performance for label 0(not recommended reviews). Therefore, below analysis will focus on label 0.

For label 0, the model train with 5 epochs has the highest precision value at 0.86. This indicates that for all predictions that the model gives, 86% of them truly belong to label 0.

In terms of recall, the model training with 4 epochs improves the most compared to the baseline. This indicates that the model correctly identifies approximately 77 % of all label 0 in the validation set.

Furthermore, for the f1-score, other than the model with 5 epochs, every fine-tuned model gains similar improvement over the baseline model(0.77). The reason behind lower improvement for model with 5 epochs(0.74) is because of it's low recall at 0.64.

Overall, fine-tuned models did improve the performance over the baseline model. The model with 4 epochs is the most balanced model to achieve good precision, recall, and F1-score. Also, model training with 5 epochs achieves higher precision than others. There is also no clear sign of overfitting on the 10 epochs model. However, it requires substantially more time for training as shown in previous section. Therefore, the manual analysis will be conducted for the model training for 4 epochs and 5 epochs.

Table 3: Classification Report of the Fine-Tuned Model with 3 Epochs

	precision	recall	f1-score	support
0	0.80	0.74	0.77	1031
1	0.97	0.98	0.97	8588
accuracy			0.95	9619
macro avg	0.89	0.86	0.87	9619
weighted avg	0.95	0.95	0.95	9619

Table 4: Classification Report of the Fine-Tuned Model with 4 Epochs

	precision	recall	f1-score	support
0	0.78	0.77	0.77	1031
1	0.97	0.97	0.97	8588
accuracy			0.95	9619
macro avg	0.87	0.87	0.87	9619
weighted avg	0.95	0.95	0.95	9619

5.4 Manual Analysis

For manual analysis, there are multiple reviews that can be seen as Scarsam. Below are some examples

Table 5: Classification Report of the Fine-Tuned Model with 5 Epochs

	precision	recall	f1-score	support
0	0.86	0.64	0.74	1031
1	0.96	0.99	0.97	8588
accuracy			0.95	9619
macro avg	0.91	0.81	0.85	9619
weighted avg	0.95	0.95	0.95	9619

Table 6: Classification Report of the Fine-Tuned Model with 10 Epochs

	precision	recall	f1-score	support
0	0.80	0.73	0.77	1031
1	0.97	0.98	0.97	8588
accuracy			0.95	9619
macro avg	0.89	0.86	0.87	9619
weighted avg	0.95	0.95	0.95	9619

with the original recommendation status and the review text.

1. Negative Review:*this game is about as entertaining as watching paint dry*
2. Negative Review:*i loved the game, mainly because i've always wanted aids.*
3. Negative Review:*it's just awesome.*
4. Negative Review:*this is my favorite game*
5. Postive Review:*i dont like playing this game*
6. Postive Review:*no*

In Table7, The number and the percentage of misclassified reviews due to sarcasm are presented. The percentage of the baseline model is lower than that of the fine-tuned model. However, this is due to the amount of misclassified reviews from the baseline model being significantly higher than the fine-tuned version. By comparing the number of how many sarcastic reviews each model misclassified, there is no improvement from the baseline model.

Table 7: manual Analysis on Sarcasm

Model	Sarcasm counts	Sarcasm Percentage
Baseline	33	3.20%
Fine-Tuned 4 epochs	46	9.89%
Fine-Tuned 5 epochs	34	7.31%

6 Discussion

Comparing the baseline and fine-tuned model performance, there is a significant improvement in

the not-recommend reviews. This is improved by the fact that the model is trained on this specific dataset and can better capture the nature of the not-recommend reviews. Thus improving the accuracy of the imbalanced data.

As the manual analysis section suggests, using this specific Steam Review Dataset to fine-tune the RoBERTa model does not allow it to capture sarcastic reviews better than the baseline model. The count of sarcastic reviews within misclassified reviews is higher in the fine-tuned models compared to the baseline. Specifically, the misclassified reviews from the baseline model contain 33 sarcastic reviews, while the fine-tuned model training with 4 epochs has 46 reviews, and the model with 5 epochs has 34 reviews. One hypothesis for this result is that the overall content in the dataset is non-sarcastic. Therefore, more in-depth fine-tuning or another model shall be investigated for this type of application. The other suggestion for future work will also be to observe the raw dataset first to gain insight into the percentage of sarcastic reviews within the dataset. Furthermore, a more capable machine is recommended to train on and validate using the entire dataset.

7 Conclusion

To sum up, four RoBERTa models have been fine-tuned using the Steam Reviews Dataset 2021. The RoBERTa family model is a highly capable model for sentiment analysis on video game reviews. By fine-tuning for the downstream task such as the Steam game review dataset, the fine-tuned model achieves a more balanced performance to counter the imbalance label in the validation set compared to the baseline model. With training with 4 epochs, the model reached 0.97 on all precision, recall, and f1-score on recommended review and closed to 0.77 on all metrics for non-recommended review. Moreover, from training results, even using small epochs of 3 epochs, the fine-tuned model can achieve 96% accuracy on the training set. In terms of analysis of sarcasm, Fine-Tuning RoBERTa does not gain improvement for this type of application, and future work is needed to explore this task.

8 AI assistant clarification

For this project, codes for setting up the training of fine-tuned RoBERTa and the evaluation function are created with the assistance of CHATGPT. Also, Grammarly and CHATGPT are used to check grammar and rephrase after writing.

References

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Haibo He and Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marko M. 2021. Steam reviews dataset 2021. https://www.kaggle.com/datasets/najzeko/steam-reviews-2021/. Accessed on March 3, 2024.

Navonil Majumder, Soujanya Poria, Haiyun Peng, Niyati Chhaya, Erik Cambria, and Alexander Gelbukh. 2019. Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems*, 34(3):38–43.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jie Ying Tan, Andy Sai Kit Chow, and Chi Wee Tan. 2021. Sentiment analysis on game reviews: A comparative study of machine learning approaches. In *International Conference on Digital Transformation and Applications (ICDXA)*, volume 25, page 26.

X. Xia, Y. Huang, and Y. Zhang. 2023. Multi-task learning for game review classification with emotion and sarcasm detection. In *2023 8th International Conference on Data Science in Cyberspace (DSC)*, pages 159–165, Los Alamitos, CA, USA. IEEE Computer Society.