Analysing sentiment in professional GPU reviews and user comments

Floris Rossel
floris.rossel@ru.nl

ABSTRACT

Sentiment analysis with language models like BERT allows us to automatically assign sentiment scores to texts. We investigated using a pre-trained multilingual BERT-based model to analyse sentiment in GPU reviews on the Dutch technology website Tweakers. Our results show a range of scores between 1 and 4 stars out of 5 for the review sentiments, with most GPUs scoring between 2 and 4 stars. We also analysed sentiment of user comments that belong to these reviews, but find no correlation between review and comment sentiment.

1 INTRODUCTION

The appearance of neural language models has received much attention in the field of Natural Language Processing in the last couple of years. These language models have been applied for many use cases, such as text prediction and classification. What makes a transformer-based language model like BERT [2] very useful is the ability to fine-tune the model for specific classification tasks, one of which is sentiment analysis. In this paradigm, texts are classified as being positive or negative in sentiment or are given a discrete score such as a star rating.

In this paper, we use a pre-trained multilingual BERT-based model from 'NLP Town' [6] to analyse the sentiment of professional graphics card reviews and user comments on the Dutch technology website Tweakers. We also show to what degree the overall sentiment of user comments align with the reviews.

We also perform a validation analysis where we manually label the sentiment of comments and compare this to the scores assigned by BERT.

2 RELATED WORK

Fang et al. (2015) [4] did a similar analysis on Amazon product reviews, categorizing sentiment polarity on both the review-level and the sentence-level. Their work involves a pipeline with 3 phases, where intermediate features are created. In the final phase, the sentence level and review level sentiment polarity experiments are performed. As opposed to their work, we use a much simpler pipeline where the transformer directly outputs a sentiment score.

In our work, we analyse Dutch reviews and comments. Several Dutch BERT-based language models have been developed for the Dutch language. One of them is RobBERT, a RoBERTa-based model trained on Dutch texts (Delobelle et al., 2020 [1]). One of the tasks it was trained on was sentiment analysis. Their model outputs positive or negative labels to texts that describe the sentiment. We decided not to use this model but opt for the multilingual 'NLP Town' model instead, because this model gives more fine-grained scores. We think that extracting binary labels for our research question is not very informative, since GPU reviews are often nuanced. A wider sentiment score also lends itself better to a correlation analysis.

3 APPROACH

3.1 Data

For this research we will limit our analysis to the reviews of 46 graphics cards models from the companies AMD and Nvidia that recently entered the market in the consumer desktop segment, starting from the year 2016. The products we looked at all fall into the Radeon or GeForce naming scheme, which are the companies' main consumer product categories. The reviews are also limited to a subset of GPU models in the medium to high-end price segment, as these are the most popular and this is what Tweakers mainly focuses on. Inexpensive low-end graphics card models are rarely reviewed by Tweakers, so these are not included. The manufacturer's suggested retail prices of the GPUs that we looked at ranged from 109 all the way to 1599 $ at the time the product was launched into the market. We collected the reviews and user comments for the GPUs that fall into our selection of GPUs that were available at the time of writing. The full list of GPU models incorporated in our research is included in Table 1.

There are a small number of GPUs are combined in the same reviews. In all cases, these are very similar products that are next to each other in the product stack of their generation. The concerning models are marked by asterisks in the results. Because these models share the same reviews and therefore the same comments, the scores will be the same. The shared results prevents us from assigning the sentiments to a specific model, but we thought it useful nevertheless to give an indication for both products.

For every GPU we manually download the HTML documents from the review's web page. These HTML files include the user comments, which is often split into multiple web pages when the number of comments exceeds a limit.

Because of the input size limitation of the language model, we limited our analysis to the conclusion segment of the review. We argue that the conclusion section contains a balanced overview of the positive and negative aspects of a product, and is therefore well suited for extracting a sentiment score on a 5-star scale.

For each review, we extract all user comments that are written under the article, including reactions to comments. The number of comments for all reviews ranges from 74 to 602, with an average of 212. These comments often contain URLs and quotes of other comments, which we remove in our data pre-processing pipeline.

3.2 Data pre-processing

In order to get the input for the BERT language model, we use the Python package BeautifulSoup to extract specific HTML elements that contain the relevant text. For our model input, we include

TxMM'22-'23, Radboud University, Nijmegen, Netherlands

Table 1: Reviews of GPU models included in our research

Nvidia	AMD
RTX 4090, RTX 4080, RTX 3090 Ti, RTX 3090 RTX 3080 Ti, RTX 3080, RTX 3070 Ti, RTX 3070, RTX 3060 Ti, RTX 3060, RTX 3050 , RTX 2080 Ti, RTX 2080 , RTX 2070, RTX 2060, GTX 1660 Ti, GTX 1660 Super, GTX 1660, GTX 1650, GTX 1080 Ti, GTX 1080, GTX 1070 Ti, GTX 1070, GTX 1060, GTX 1050 Ti, GTX 1050	RX 7900 XTX, RX 7900 XT, RX 6900 XT, RX 6800 XT, RX 6800, RX 6700 XT, RX 6600 XT, RX 6600, RX 6500 XT, RX 5700 XT, RX 5700, RX 5600 XT, RX 590, RX 580, RX 570, RX 480, RX 470, Radeon VII, RX Vega Liquid, RX Vega 64, RX Vega 56

all subtitles in the HTML headers, as these often contain value statements about the product. All non-text elements are excluded. The raw HTML is then converted to the text that is shown on the web page by removing the HTML formatting elements.

For the comment extraction, we extract the text in each comment block and store this as a separate object. Tweakers allows users to quote (parts of) other other comments, which are put into a quote block. We remove these blocks from the comment, as they are not written by the user that made this comment. Hyperlinks that start with 'http' are also removed using a regular expression.

3.3 Model

In this paper, we use a multilingual BERT-based model developed by the "NLP Town" research team [6], available in the HuggingFace Transformer Library [8]. This model is fine-tuned to predict star ratings of product reviews in 6 languages (English, Dutch, German, French, Spanish and Italian), and showed moderately good accuracy for classifying reviews on a 1 to 5-star scale in the Dutch domain. For the exact prediction of the number of stars, the model got a 57% accuracy, while it had a 93% accuracy when a difference of 1 star was tolerated.

3.4 Sentiment analysis

In order to relate review and comment sentiments, we will first run BERT on every review and every comment, resulting in a score for each. Since there are multiple comments for every review, we will aggregate the scores by taking the average. Each GPU is thereby given a review score and comment score average. In this work, we will give every comment the same weight, even though not all comments necessarily contain opinions about the product. We think that the distribution of comments 'goals' stays somewhat consistent for different reviews such that as we average comment sentiment scores, the result is still informative of user sentiment.

Ideally, we would like to input the full text of each review and comment into the BERT model. However, this is not always possible as this model has a limited input size of 512 tokens. There are several ways to handle this problem. Chi Sun et al. (2019) [7] looked at three truncation methods to limit the sequence such that it fits. Since the beginning and end of a text often contain the most important points, they tested using the head, tail and a combination of the head and tail of the text for classification. The latter consists of a 128-382 token split such that the tail is the largest. Their work shows that using the head and tail gives the best results, closely followed by the tail-only method.

We use a slightly modified approach, where we initially balance the head and tail size to 255 tokens each and then move the truncation points to the periods at the end of sentences such that sentence context is kept intact. Any leftover tokens from the head are then given to the tail to maximize the amount of text used by the model. We argue that using a more balanced split is better suited for the conclusion sections of a review, since it is already more concise than a generic text, and therefore has a more uniform distribution of value statements about the product. Comments that exceed the token limit are quite rare, so for these, we did not alter the splitting ratio.

When all scores are generated, we calculate the Pearson correlation coefficient between average comment scores and review scores. This will tell whether review sentiment has an effect on comment sentiment.

3.5 Comment sentiment validation

In order to validate whether the model is able to accurately assign sentiment scores, we will conduct a small analysis where we manually label comments with a sentiment score and compare these with the generated scores from the model. From all comments of all reviews, we randomly sample 40 comments and give these a 1-5 star score each. The model-assigned scores are not shown to prevent this from affecting the manual labelling. The score differences are then counted and will inform us of the accuracy of the model. These differences are merely an indication of model validity, since the sample size is very small and the manual labels are based on a single person.

4 RESULTS AND ANALYSIS

Table 3 shows the sentiment scores of the reviews and the average comment scores. It is notable that no GPU review gets a maximum score of 5 stars, while only one scores 1 star.

Figure 1 shows how review scores and average comment scores are distributed. We can see that a 2-star rating is most common, while 3 and 4-star ratings are less so. The sentiment scores for the reviews and comments showed a Pearson's correlation coefficient of 0.133, but this is statistically insignificant (p= .373).

Our results indicate that there might be a small correlation between review and comment sentiment, but our work was not able to find one that is statistically significant. Sentiment expressed by user comments do not seem to be affected by review sentiment according to the BERT model we used.

4.1 Comment sentiment validation

Our manual validation of comment sentiments showed that the model was not much different in assigning scores than we did. Out of the 40 comments we sampled and manually rated, BERT gives the same rating in 16 (40%) cases (Table 2). If we tolerate a difference of 1 star, the model gives an adequate rating in 35 (87.5%) cases.

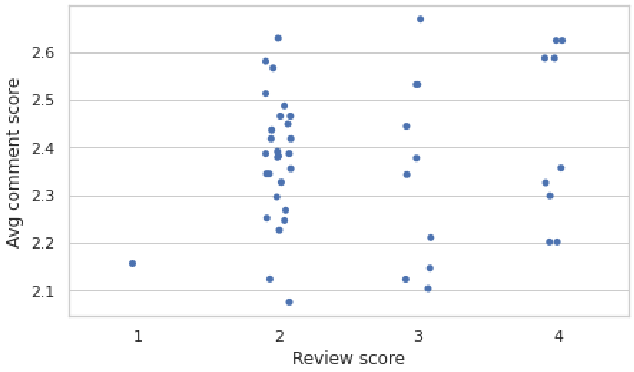

Figure 1: Review and average comment scores

Table 2: Score differences between manual and automatic sentiment rating ($difference = manual - automatic$**)**

Score difference	4	3	2	1	0	-1	-2	-3	-4
Count	0	0	5	9	16	10	0	0	0

It is important to note that the manual ratings are given by one person, and the sample size is not very big. However, overall we find no reason to assume that the model is inadequate for analysing sentiment of comments.

5 DISCUSSION

Our work shows how sentiment analysis using the BERT language model from NLP Town can be used to automatically assign sentiment scores to product reviews and comments. However, the model does not produce fine-grained scores, which prevents the differentiation of many graphics card reviews in their sentiment. A 5-star scale can suffice if our goal is to find products that are very badly or highly rated. For products with a lot of nuance regarding their quality, as is the case with most GPUs we looked at, the results provide less insight into their relative sentiments due to the small range in scoring.

One result that stands out from our analysis is the 1-star score for the review of the RTX 4080. This graphics card has received a lot of criticism in the media regarding the price point relative to its performance. Additionally, controversy arose regarding a second lower-performance version that carried the same name, which is a move that people can see as misleading. We also observe that the average comment sentiment scores are quite low for this graphics card compared to other cards, although it is not the lowest. However, overall we find insufficient evidence that comment sentiment is correlated to review sentiment.

We propose a few reasons why comment sentiment is not affected by review sentiment. For one, the model was trained on product reviews specifically, and not on other domains like comments. The definition of sentiment in the training data is specifically a star rating on a 5-point scale about the product in question. This is not necessarily generalizable to the sentiment that people express in comments, since they are not reviews. We argue that there is still a lot of overlap in the domain we investigated. The comments under the reviews often reflect the sentiment that users on the website have regarding the GPU in question, but they can also reflect other things like review methodology, competitor GPU models, or other things that are off-topic. To solve this problem, we need to increase the granularity of the analysis such that only the relevant statements about the product are considered. Do et al. (2019) [3] compared over 40 models for aspect-based sentiment analysis by their performance, many of which are based on deep learning methods. These models often look at semantic and syntactic textual information to extract sentiments about a certain target, which can be a product or concept. Future work can use these findings to select the best models and improve the analysis in our research.

Then there is yet another problem with connecting comments and consumer sentiment, namely that online ratings do not necessarily reflect population perception. Users that comment on Tweakers' reviews are a very small subset of potential consumers of these graphics cards but are not representative. These users are probably more technically minded and enthusiastic about these products than the average consumer in this market. Be also believe there exists a bias for people to comment on their viewpoint when they are negative, which skews the sentiment scores.

The question of online ratings versus population perspective was examined in a study by Gao et al. (2015) [5]. They found a positive correlation between online ratings and population opinions of physicians, which suggests that it is useful to look at online expressions of opinion when the goal is to know more about population perspectives.

There are many shortcomings in our research that could be addressed in future works. A simple aspect that can be improved is the sample size, by including more product reviews, perhaps also in other product categories. In order to get a better view of comment sentiments towards the products, one could use algorithms to classify which comments express an opinion about the product, and which ones do not. We think that omitting the irrelevant comments will result in the biggest effect on the correlation with review sentiment.

One could also improve the analysis for the individual comments by making the language model pay attention to only the most relevant parts that talk about the product. Especially in comments that compare competing products will this be beneficial, because our current model does not differentiate the opinion of different products.

6 CONCLUSION

We demonstrated that the pre-trained BERT-based model from the authors at "NLP Town" is well suited for analysing the sentiment of professional graphics card reviews. Our manual validation further shows that this also works for user comments. We constructed a list of scores on a 1 to 5-star scale for reviews of 47 recent graphics card models from the manufacturers Nvidia and AMD. We also show how average comment sentiment is not significantly correlated to review sentiment. Future work that uses more sophisticated sentiment analysis methods, like aspect-based approaches, may find positive results.

REFERENCES

[1] Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. *CoRR* abs/2001.06286 (2020). arXiv:2001.06286 https://arxiv.org/abs/2001.06286

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[3] Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review. *Expert Systems with Applications* 118 (2019), 272–299. https://doi.org/10.1016/j.eswa.2018.10.003

[4] Xing Fang and Justin Zhan. 2015. Sentiment analysis using product review data. *Journal of Big Data* 2, 1 (2015), 1–14.

[5] Guodong (Gordon) Gao, Brad N. Greenwood, Ritu Agarwal, and Jeffrey S. McCullough. 2015. Vocal Minority and Silent Majority: How Do Online Ratings Reflect Population Perceptions of Quality. *MIS Quarterly* 39, 3 (2015), 565–590. https://www.jstor.org/stable/26629621

[6] J. Leys and Y. Peirsman. 2020. Bert-base-multilingual-uncased-sentiment. Retrieved April 18, 2022 from https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

[7] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to Fine-Tune BERT for Text Classification? *CoRR* abs/1905.05583 (2019). arXiv:1905.05583 http://arxiv.org/abs/1905.05583

[8] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR* abs/1910.03771 (2019). arXiv:1910.03771 http://arxiv.org/abs/1910.03771

7 APPENDIX

Table 3: Review and comment sentiment ratings for Nvidia GeForce and AMD Radeon GPU reviews

GPU	Stars	Average comment score	#comments
RTX 4090	2	2.437	307
RTX 4080	1	2.159	270
RTX 3090 Ti	2	2.296	152
RTX 3090	2	2.382	246
RTX 3080 Ti	2	2.076	159
RTX 3080	2	2.630	602
RTX 3070 Ti	2	2.379	116
RTX 3070	3	2.378	225
RTX 3060 Ti	4	2.357	168
RTX 3060	2	2.226	137
RTX 3050	2	2.567	127
RTX 2080 Ti*	2	2.345	342
RTX 2080*	2	2.345	342
RTX 2070	3	2.147	191
RTX 2060	3	2.123	154
GTX 1660 Ti	2	2.247	150
GTX 1660 Super	2	2.581	74
GTX 1660	3	2.444	117
GTX 1650	2	2.327	104
GTX 1080 Ti	3	2.669	130
GTX 1080	4	2.202	238
GTX 1070 Ti	2	2.251	139
GTX 1070	4	2.202	238
GTX 1060	4	2.326	267
GTX 1050 Ti*	4	2.624	205
GTX 1050*	4	2.624	205
RX 7900 XTX*	2	2.466	350
RX 7900 XT*	2	2.466	350
RX 6900 XT	4	2.298	181
RX 6800* XT	2	2.418	404
RX 6800*	2	2.418	404
RX 6700 XT	2	2.392	166
RX 6600 XT	2	2.268	179
RX 6600	3	2.211	147
RX 6500 XT	3	2.104	164
RX 5700 XT*	3	2.532	171
RX 5700*	3	2.532	171
RX 5600 XT	2	2.356	90
Radeon VII	2	2.513	187
RX Vega Liquid	2	2.124	186
RX Vega 64*	2	2.387	235
RX Vega 56*	2	2.387	235
RX 590	2	2.487	158
RX 580*	4	2.588	114
RX 570*	4	2.588	114
RX 480	2	2.449	403
RX 470	3	2.343	169