H4rdc0r3 vs Casuals: Examining Video Game Reviews and Genres Through Classification and Aspect Extraction

Shujing Dong, Vidhu Nath, Michael Shum
{sjdong, vid.nath, mikeshum} @berkeley.edu

Abstract

The current state of natural language processing (NLP) has yielded many models that train on consumer product reviews for providing more refined recommendations and enhancing user feedback. While the primary language of product reviews follows this paradigm, video game players have developed their own styles of communication with slang, abbreviations, and text substitution that have resulted in a corpus unique to these gamers and not fully similar with many other product review data. In this paper we build on previous literature done on product reviews and video game reviews to examine the effectiveness of NLP classification models on video game reviews. We also build on work done in aspect-based sentiment analysis (ABSA) and determine how effective existing techniques in this field are when applied to the video game review text. We conclude with investigating both classification and ABSA performance on three video game review dataset by different genres to see if there are differences in performance from the underlying corpus. We find that the state-of-the-art classification models have high accuracy at determining recommended games; that aspect-based extraction yields meaningful sentiment analysis around the extracted aspects; and that for both classification and ABSA there is a clear difference between using data from games played by casual vs hardcore gamers.

1 Introduction

The video game industry is now worth about $159.3 billion USD in 2020 according to Statista, more than the music and movie industries combined (Gough, 2020). An estimated 75% of all households in the US have someone who plays video games. Despite becoming mainstream, people who play video games have developed their own video game vernacular, derived from common speech but morphed into a composition of shorthand and character replacement that is suited to being typed and used in the context of gaming. Sarcasm and "trolling" is commonplace, as is a lot of game-specific jargon.

One example form of expression is through "leetspeak", where letters in words are replaced by numbers to bypass profanity filters in online chat forums and game threads - the term itself comes from "leet" which is short for "elite", and is often stylized as "1337". Many words have also been created that are specific to gaming, such as hardcore players calling newer gamers *n00bs* (newbies, or new players) or *hax0rz* (hackers, or people being accused of cheating in online games). This unique vocabulary utilized by gamers has created an entirely different subculture with its own corpus, which has been studied in papers such as one by Bawa (2018) that examined players in different groups within massively-multiplayer online role-playing games (MMORPGs).

The *leetspeak* approach to player language and gamer-specific slang is greatly contrasted from that of other forms of entertainment such as movies and television, where the linguistic structure of the text is relatively consistent with natural descriptive language (Yin et al., 2018). This presents interesting challenges to natural language processing models. Another issue is that words used can vary greatly depending on the game, which could impart an uncertain, potentially negative impact on model performance.

In this paper, we explore three different hypotheses related to NLP and their application on a video game review corpus.

Hypothesis 1 (H.1.): State-of-the-art NLP models can outperform baseline models in text classification tasks on a Game Review corpus.

Hypothesis 2 (H.2.): Aspect-based Sentiment-Analysis (ABSA) with unsupervised attention-based aspect extraction models are applicable to a Game Review corpus.

Hypothesis 3 (H.3.): The type of game (hardcore vs casual) of the game review will provide meaningful

differences in both the recommendation classification of the review as well as in aspect sentiments from unsupervised ABSA.

For our first hypothesis (H.1), we evaluate the predictive performance of a range of different NLP algorithms when used on a video game corpus of text. By starting with simpler and more established models, such as Bag of Words with XGBoost, then graduating to more state-of-the-art algorithms like BERT, we examine the comparative benefits the varying level of algorithms deliver when training on a dataset inherently different from most text models, similar to how our reference paper Ni et. al (2019) proceeded. Based on previous background research detailed in the next section, we hypothesize that the advanced algorithms like LSTM and BERT will perform consistently better than more baseline algorithms like BoW or CNNs despite training on a relatively segregated corpus from more common applications.

Second, we implement an unsupervised sentiment analysis model utilizing aspect extraction to examine the applicability of the technique with this dataset, specifically to see if aspect extraction can be reasonably conducted unsupervised for this corpus. Our second hypothesis (H.2) is that with some tweaks, this type of analysis would be feasible and potentially effective on a video game corpus with differences in the positivity/negativity of the sentiment clear between aspects.

Finally, we approach the first two hypothesis using three sets of data: “casual” games, “hardcore” games, and combined. “Casual” games (referred onward as casual) that we collected focus on teamwork, communication, and are made to be more easily learned for a variety of players. “Hardcore” games (referred onward as hardcore) that we picked comprise primarily of first-person shooting (FPS) games since they are more associated with toxic or unapproachable communities of players, even within gaming overall (Smith, 2020). Our last hypothesis (H.3) is that these datasets will show differences in our classification and ABSA tasks that reflect the difference in genre due to inherent linguistic differences even with gaming as a whole (Bawa, 2018). Casual game review data will act as a baseline vs a hardcore game review dataset.

2 Background

The work we seek to accomplish with this investigation is broken into two main tasks: the classification of the review data as recommendations and the extraction of aspects from the review for sentiment analysis. Both approaches have a large corpus of literature behind them proving their value, though it is more limited when it comes specifically to video games and their reviews.

Ni et al. (2019) is a closely related paper we used for reference due to the overarching design including classification and aspect extraction. They sought to extract fine-grained aspects for providing justifications behind reviews to increase transparency and consumer understanding. Their work utilizes a data generative approach of annotations that are then passed into a classification pipeline with bag-of-words (BOW), a convolutional neural network (CNN), long short-term memory (LSTM) model, and BERT. The output is passed into an aspect conditional masked language model (ACMLM) that generates the justifications for their reviews. Their initial goal was different from ours, however, so we leverage their work up until it becomes different due to the supervised aspect extraction and the use of the ACMLM.

Despite video game reviews not being studied as thoroughly as other product domains, there still exists meaningful research that we leverage to understand this domain. A recent report by Viera et al. (2019) aimed to determine user acceptance of games based on reviews by implementing a CNN – specifically a GloVe DCNN. They apply the DCNN to classify sentiments as “positive”, “neutral”, and “negative”, which we do as well for our aspect-extraction task. We leverage their application of the DCNN to make a similar CNN, but their approach ends with the DCNN whereas we implement further classifiers and aim for “recommended” or “not recommended” for our classification task. Additional work we relied on has expanded sentiment classification with newer models such as attention-based LSTMs that include aspects (Wang, 2016) and sentiment classification using BERT with fine-grained sentiment extraction and processing (Munikar et al., 2019). Other video game research we investigated included user satisfaction text analysis (Wang et al., 2020).

Aspect extraction has also been widely researched as part of sentiment analysis, specifically for reviews. Li et al. (2019) provided significant value with their work into applying BERT for an end-to-end (E2E) ABSA model with two datasets, though they use a supervised model whereas we use an unsupervised model. Additional research by Tulkens et al. (2020) allowed us to leverage their unsupervised model for aspect extraction which was a challenge we initially sought to avoid. Hutto and Gilbert (2015) provided an overview of a rule-based model for sentiment extraction that we utilize as part of our ABSA pipeline. Along with some of our video game

research described above, we additionally relied on Zagal et al. (2009) for understanding what aspects of a game would be worth utilizing for our sentiment extraction.

3 Method

3.1 Corpus Generation

To generate our video game review text corpus, we use a pre-coded web crawler called *STEAM crawler* that we adapted for our own purposes, originally provided by the user *aesuli* on GitHub (Esuli, 2020). It scrapes game reviews written by users on the online video game store Steam, which meant that our data was written by gamers for a gaming audience. New reviews are written daily for tens of thousands of games, resulting in a fresh and constantly updated corpus of review text.

As mentioned in the introduction, we focus on the casual and hardcore genres for our video game review selection. From Barbieri (2016) we see that gamers develop their own communication and sub-culture based on their groupings, so we hypothesize that the differences in genre will be apparent due to the groupings of players that consume those games (H.3.). Casual games generally are more mainstream thus more accessible to a wider audience. They also tend to require teamwork and communication to achieve goals; we surmise that these qualities of the game reflect the player – and thus the review quality. We have similar expectations of the hardcore genre, represented exclusively by FPS games for our investigation due to the media attention, dedicated fanbases, and use of direct confrontation as core game mechanics. Like for casual games, we expect FPS games to have reviews which might more closely reflect their often-divided player base. Note that going forward, we interchangeably use 'hardcore' and 'FPS'.

Specific games we choose are as follows: for casual, we select *Among Us, Phasmophobia, The Jackbox Part Pack 7, Fall Guys, FIFA 2021, Football Manager 2020,* and *NBA 2K20*. These games include a mix of team-objective base gameplay as well as cooperative sports games. For FPS games, we decided on *Counterstrike: Global Offensive, PlayerUnknown's Battlegrounds, Destiny 2, and Warframe*. These are popular FPS games with thousands of active users daily.

3.2 Data Preprocessing

After collecting review data for the games previously described, we create a text pre-processing pipeline that consists of combining and cleaning. All the data are appended to the genre-specific sets – for casual and hardcore – as well one combined dataset with all the data. We pass these to a basic text processing suite which changed case, removed unwanted characters like 'https' from linked webpages, removed punctuation and numbers, and removed words that were in a set of 'shortwords' – those which are less than two characters in length. These steps improved on inconsistencies in the data, but does not account for typos, misspelled words, or other random strings.

Following the text preprocessing, we apply manual tokenization to the text strings to allow us more control over the review data contents since we expect uncommon or non-conforming text due to the linguistic idiosyncrasies of gamer language. Part of this step included removing symbolic language like *emojis* which are common in these reviews. Felbo et al. (2017) showed that emojis and other symbolic characters have presented richer representations of text and nuance, but choose to remove them for simplicity in processing and analysis.

3.3 Exploratory Data Analysis

We conduct exploratory data analysis (EDA) briefly to gain an intuition about the nature of our data. Overall, we interface with over 500,000 reviews; there is a significant skew towards FPS games due to some limitations with accessing permissions to some games on Steam which affected casual games more so. While we do not focus on numerical analysis or regression with this research, we still find some relationships worth noting.

As shown in Table 1, FPS games had much larger average play time logged than casual games (309 vs 50 hours) as well as somewhat longer review lengths (10.3 vs 8.9 words). Casual games had higher average user scores than FPS games (0.84 vs 0.51 recommended on a scale of -1 to 1). There might be several explanations for these results: casual games might have fewer reviews than FPS due to consistently high recommended values – thus a lower need to convince a new gamer to play. Conversely, higher play times in a game might result in longer reviews due to more understanding, as we see and predict with FPS games. However, we do not examine these results further in our research; they are more useful in gaining a high-level understanding of the data.

Dataset	Number of Reviews	Average Review Length	Average Hours Played	Average Recommended
FPS	374,266	10.34 ± 12.97	309.80 ± 262.21	0.51 ± 0.86
Casual	180,266	8.92 ± 10.05	50.44 ± 95.93	0.84 ± 0.548
Combined	553,767	9.67 ± 11.80	223.64 ± 252.84	0.62 ± 0.78

Table 1: Dataset Statistics

3.4 Modeling Recommendation Classification

The first major component of our approach is the *classification* task for game recommendations. We approach this task as Ni et al. (2019) did and apply a suite of established NLP algorithms as well as more modern approaches. First we split our data in a 80:20 ratio for training and testing. The four models we train and test on are: XGBoost using Bag-of-Words (BOW) for sentence features, a convolutional neural network with global max pooling, LSTM with global max pooling, and BERT (see model structures in Appendix). We decide to use BOW and CNN since they are common and have years of use, representing the first major NLP-focused algorithms following machine learning. LSTM and BERT are models that have shown to be exceptionally useful and generalizable, but are newer and only now being widely accepted, meaning that we are able to test their expected performance improvements against data that they may not have trained on consistently.

3.4 Modeling Unsupervised Aspect-based Sentiment-Analysis

The second major component of our approach was to enact an unsupervised model for aspect-based sentiment analysis. As described in the background literature, we utilize Tulkens et al. (2020) and their way of extracting aspect terms to assign aspect label to the reviews. This entire approach is done in two steps, first manually selecting aspect labels and automatic assigning of the aspects to reviews, and second, entailing sentiment score calculations for each review and reference the score with the assigned aspect.

We first parse out each token and assign it a part-of-speech per the **spaCY** library, which was referenced by Honnibal and Montani (2017). After that we use an attention mechanism to create a matrix that operates on an aspect to get a weighted summary for a sentence. This can be described by first getting an attention value *att* for a given aspect *a* and a matrix S:

$$att = softmax(aS)$$

$$d = \sum_i att_i S_i$$

Where we calculate a weighted sentence summary *d*. These sentence summaries are applied to create a contrastive attention aspect extraction model as shown in Figure 1 from an example provided by Tulkens et al. (2020), the output of which are sent to a method that extracts sentiment score for each review based on the work of Hutto et al. (2014), which was originally purposed to extract aspect sentiment from social media posts. The process we use to extract aspects is based on the work of Zagal et al. (2009), from where we learn to apply the strongest resulting aspects of a game as it's more meaningful among the reviews. We conclude the ABSA pipeline by merging sentiment score and aspects for each review into a single value for each genre and dataset, allowing us direct comparisons for the impact of each aspect for an example game in each genre.

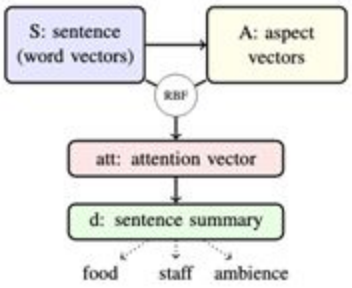

Figure 1: General Contrastive Model using Aspects(from Tulkens &Cranenburgh(2020) **)**

4 Results

4.1 Classification Model Performance

Table 2 shows the sets of outcomes of the classification models on our data. We run our models on both the genre-specific data as well as combined to provide evidence for H.3 – that different genres will have different results due to the style of gameplay and players. Across the models there is reasonably high accuracy, but the difference between the models and the datasets becomes clear.

Our first hypothesis H.1. is confirmed by examining the accuracy for each dataset across the models. Consistently, XGBoost does worst and BERT performs best, with increased accuracy between

CNNs and LSTM as well. This is true for each genre and our combined dataset, so we believe we have been able to validate H.1 as well the classification task component for H.3. There is a noticeable jump in the accuracy between genres at even the earlier models – 0.85 for XGBoost on casual games vs 0.79 for XGBoost in FPS games - which predictably evens out in the combined dataset. This outcome is consistent across the precision and F1 metrics as well, indicating to us that we have proven H.1. and part of H.3. In particular, the BERT model was able to get high Precision(0.92+) across all genres, indicating its advantage in text classification. (Note: CNN and LSTM are trained with 10 epochs, batch size=128, BERT is trained with 10 epochs, batch size =512).

Model	*Accuracy*	*Precision*	*F1*
Casual Games Dataset			
XGBoost	0.85337	0.79200	0.85312
CNN	0.86420	0.82516	0.86402
LSTM	0.91687	0.89174	0.91681
BERT	0.92140	0.94300	0.91910
FPS Games Dataset			
XGBoost	0.78671	0.71582	0.78574
CNN	0.80400	0.73926	0.80383
LSTM	0.84966	0.79566	0.84965
BERT	0.86910	0.92870	0.85932
Combined Dataset			
XGBoost	0.79266	0.72219	0.79179
CNN	0.81763	0.77246	0.81706
LSTM	0.87405	0.82150	0.87399
BERT	0.87560	0.93540	0.86664

Table 2: Outcomes of Recommendation Classification

4.2 ABSA Model Performance

Figure 2 and Figure 3 highlight the results of the contrastive attention-based ABSA model with our data, using an example game for each genre. We successfully extracted aspects consistently across the review data for both genres. In our pipeline, we manually select aspect labels “game”, “tactic”, “design”, “feeling”, and “community” according to earlier game review research (Zaga et al., 2009); then we automatically assign aspect labels to each review with the contrastive attention algorithm. For example, if reviews contained the phrases “this has terrible gameplay” and “levels are tough to finish”, it would attribute weight to the “feeling” and “design” aspects. This outcome supports the capabilities of the unsupervised ABSA and lends credence to H.2.

Our result shows in casual games, the “game” aspect displays the strong difference in the “positive” and other for game recommendation based on the reviews out of all the aspects, whereas the “design” aspect is true for our example hardcore game. The contrastive attention-based approach we use shows that it can successfully extract meaningful aspects from games. This result shows that we have validated H.2.

Additionally, we can see significant differences between the counts of the aspects between genres. The “design” aspect is largest in hardcore games, but between casual and hardcore games it is more prevalent by almost a factor of 2; this shows that the aspect is far more important to hardcore games than casual. While there are similarities in the decreased prevalence of the “feeling” and “community” aspects between genres, there are still strong differences between the “game”, “tactic”, and “design” aspects, which is evidence of the aspect component of H.3. being validated.

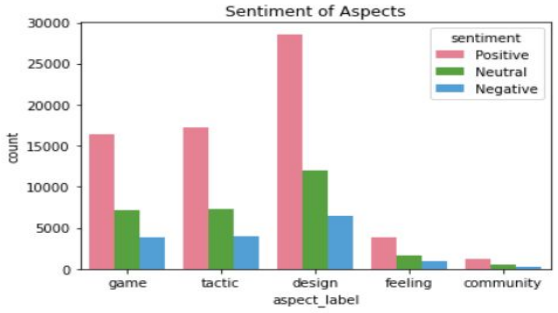

Figure 2: Aspect Extraction for a hardcore game (CS: GO)

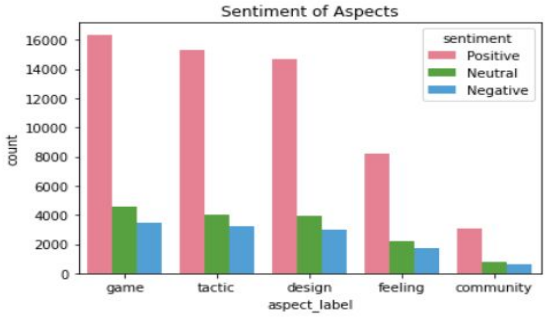

Figure 3: Aspect Extraction for a casual game (Fall Guys)

5 Conclusion and Future Work

In this paper, we examined the application of existing classification and unsupervised ABSA models on a video game review data corpus. To understand the semantic differences present in video game subculture, we input our data into a suite of established and modern classification models to see recommendation outcomes. These models were able to successfully parse the reviews and produce high-accuracy recommendations with scores consistently different across genres. We also implemented an unsupervised ABSA model that extracted five aspects from the review data and scored the sentiment of the reviews based on these aspects. Our results showed that we were successful at extracting meaningful aspects and were able to see clear differences in the counts of the aspects – thus their impact on the reviews – between genres. The results from our combined tasks validated the three hypotheses that we formulated based on previous work.

There are further enhancements we can apply in the future to improve the quality of our work. First, we might try to include more genres and more games in our data – like role-playing games (RPG) or action-adventure – and discern differences between them since there might be further divisions in subcultures based on games. We could try to account for symbols in our texts and attribute value to those symbols rather than parse them out. There is also an option for trying out conversions between English and “leetspeak”, which could allow us to test consistency of the semantics in reviews. On the evaluation side, we might try even newer classification models within transfer learning like T5 or RoBERTa and determine if gains are incremental or significant. Another enhancement could be to try more extensive aspect extraction, by using more than just 5 or selecting different aspects for different genres. Such enhancements could yield further work that advances NLP especially with video game reviews.

6 References

Christina Gough. “Video Game Market Value Worldwide 2023.” *Statista*, Statista, 28 Aug. 2020, www.statista.com/statistics/292056/video-game-market-value-worldwide/.

Papia Bawa. (2018). Massively Multiplayer Online Gamers’ Language: Argument for an M-Gamer Corpus. *The Qualitative Report*, *23*(11), 2714-2753. https://nsuworks.nova.edu/tqr/vol23/iss11/8

Fulian Yin, Yanyan Wang, Xingyi Pan, and Pei Su. "A Word Vector Based Review Vector Method for Sentiment Analysis of Movie Reviews Exploring the Applicability of the Movie Reviews." 2018 3rd International Conference on Computational Intelligence and Applications (ICCIA) (2018).

Noah Smith. "Racism, Misogyny, Death Threats: Why Can't the Booming Video-game Industry Curb Toxicity?" *The Washington Post*. WP Company, 23 July 2020.

Jianmo Ni, Jiacheng Li, and Julian Mcauley. "Justifying Recommendations Using Distantly Labeled Reviews and Fine-Grained Aspects." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019). *ACL Anthology*. Nov. 2019.

Augusto Vieira and Wladmir Brandão. “Evaluating Acceptance of Video Games Using Convolutional Neural Networks for Sentiment Analysis of User Reviews.” *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, 2019.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. "Attention-based LSTM for Aspect level Sentiment Classification." *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (2016). *ACL Anthology*. ACL, Nov. 2016.

Manish Munikar, Sushil Shakya, and Aakash Shrestha. "Fine-grained Sentiment Classification Using BERT." *2019 Artificial Intelligence for Transforming Business and Society (AITB)* 1 (2019). *ResearchGate*. Tribhuvan University, Nov. 2019.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. “Exploiting BERT for End-to-End Aspect-Based Sentiment Analysis.” *EMNLP-IJCNLP-2019*, October 2, 2019, 1–8.

Xiaohui Wang, and Dion Hoe-Lian Goh. "Components of Game Experience: An Automatic Text Analysis of Online Reviews." *Entertainment Computing* 33 (2020): 100338. *ScienceDirect*. Mar. 2020. Web.

Stéphan Tulkens, and Andreas Van Cranenburgh. "Embarrassingly Simple Unsupervised Aspect Extraction." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020). *ACL Anthology*. July 2020.

C.J. Hutto and Eric Gilbert. (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.

Jose Zagal, Amanda Ladd, and Terris Johnson. (2009). Characterizing and understanding game reviews. 215-222. 10.1145/1536513.1536553.

Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. "What Does This Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis." ACL Anthology. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), May 2016.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. "Using Millions of Emoji Occurrences to Learn Any-domain Representations for Detecting Sentiment, Emotion and Sarcasm." Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing EMNLP.SIGDAT (2017): 1615-625.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Software package.

Björn Strååt, Harko Verhagen, and Henrik Warpefelt. (2017). Probing user opinions in an indirect way: an aspect based sentiment analysis of game reviews. 1-7. 10.1145/3131085.3131121.

7 Appendix

7.1 Original Proposal

Original proposal (modified extensively):
https://docs.google.com/document/d/1_dETXpvPpZFkrmEMWKIF4DMwa_eyrAPf_I_eKC8IjU4/edit#

7.2 Classification Model structure

- **CNN**

Layer (type)	Output Shape	Param #
text_vectorization (TextVect	(None, 60)	0
embedding (Embedding)	(None, 60, 128)	25600000
conv1d (Conv1D)	(None, 60, 5)	1925
global_max_pooling1d (Global	(None, 5)	0
dense (Dense)	(None, 64)	384
dense_1 (Dense)	(None, 1)	65

Total params: 25,602,374
Trainable params: 2,374
Non-trainable params: 25,600,000

- **LSTM**

Layer (type)	Output Shape	Param #
text_vectorization (TextVect	(None, 60)	0
embedding_1 (Embedding)	(None, 60, 128)	25600000
lstm (LSTM)	(None, 60, 60)	45360
global_max_pooling1d (Global	(None, 60)	0
dropout (Dropout)	(None, 60)	0
dense_2 (Dense)	(None, 50)	3050
dropout_1 (Dropout)	(None, 50)	0
dense_3 (Dense)	(None, 1)	51

Total params: 25,648,461
Trainable params: 48,461
Non-trainable params: 25,600,000

- **BERT**

Layer (type)	Output Shape	Param #
keras_layer_5 (KerasLayer)	(None, 50)	48190600
dense_15 (Dense)	(None, 16)	816
dense_16 (Dense)	(None, 1)	17

Total params: 48,191,433
Trainable params: 48,191,433
Non-trainable params: 0

7.3 Codes for game review data collection and aspect extraction

https://github.com/aesuli/steam-crawler
https://github.com/nijianmo/recsys_justification
https://github.com/clips/cat