# GameWiki: Aspect Extraction for Video Games

Stanford CS224N Custom Project

**Sirisha Akkanapragada**
Department of Computer Science
Stanford University
sakkanap@stanford.edu

## Abstract

In this project we aim to **1)** Predict if a game review is helpful or not **2)** Extract aspects from helpful game reviews. This is helpful for gamers in identifying the most interesting aspects as well as the most disliked aspects of a game before purchasing it. The text classification task of identifying helpful reviews is implemented using a ULMFit model[2]. Past works use topic models to extract aspects, but these extracted aspects are not highly coherent. This project aims to use an unsupervised neural attention model for aspect extraction described in [1] and apply gaming domain knowledge like genre to improve the aspect extraction. The neural attention model improves coherence since it uses neural word embeddings which consider the distribution of word co-occurrences. Further, the interpretability of the aspects have been improved by splitting the dataset by genre. It was able to predict more granular aspects particular to each game genre as opposed to aspects generated from the entire gaming dataset.

Mentor: Andrew Wang

## 1 Introduction

Video gaming industry is estimated to have gained a revenue of $179.7 billion dollars in 2020 due to the pandemic[8]. With the launch of next generation gaming consoles by PlayStation, Xbox and speculated release of Nintendo Switch Pro, the industry is likely to grow much further in the years to come. Also, with the launch of next generation consoles with new features such as haptic feedback and an improved VR experience, game studios are beginning to release more creative games to immerse the gamers into the futuristic gaming worlds. There is an estimated 3,843 PlayStation games[9] on the platform that is consisting of PlayStation exclusives and cross platform games. We can estimate that there could be ~10000 games across the three major gaming console platforms - PlayStation, Xbox and Nintendo.

For a gamer, it is very hard to identify the games to buy based on their interests. While recommendation systems on all these gaming consoles suggest games that the user likes, it still doesn't mention the key aspects of the game and what makes it unique. Gaming consoles also don't mention features that gamers usually dislike about a particular game. Aspect extraction is also a precursor for aspect-based sentiment analysis which can summarize the game reviews for the end users.

In this project, we aim to identify aspects from gaming reviews that are helpful. Users upvote reviews to indicate if the review is helpful or not. While aspects could be extracted from all reviews, some reviews are very short or they don't highlight any particular game aspects. So, we use a ULMFiT classifier to classify Steam reviews if they are useful or not. The trained classifier is then used to predict if Metacritic reviews are helpful or not. The predicted helpful reviews are then used for the aspect extraction task.

The task of aspect extraction is a two fold task 1) Aspects are extracted from the reviews (like gameplay, effects, music, visuals) 2) Grouping similar terms of aspects into categories that relate to a particular aspect. While there are several labeled training datasets that could be used for aspect extraction, there is no labeled dataset for gaming data. Supervised aspect extraction methods require a lot of labeled training data and it is very hard to manually label gaming data. To tackle this problem, we have used an unsupervised neural attention model for aspect extraction. Further in this project we have improved the aspect extraction task by including game genre.

## 2 Related Work

Aspect extraction is a precursor to aspect sentiment analysis. Aspect extraction can be broadly classified into three different strategies: 1) Rule-based 2) Supervised approaches 3) Unsupervised approaches.

Rule-based approaches can only extract the top words/topics, but they don't group the topics. Supervised approaches require a lot of labeled data and transfer learning cannot be done from one domain to the other. Unsupervised approaches do not require any labeled data.

In [10], Hu et al. used rule based methods like extracting nouns and noun phrases to identify aspects. Further they used Wordnet to identify synonyms and antonyms of these of these extracted aspects. These models would not work for the gaming dataset since the domain knowledge is different and there are several aspects of games that don't fall under noun phrases categories. It could extract names of games as aspects which is not ideal in our scenario.

Supervised approaches suggested using Conditional Random Field(CRF) and deep learning based approaches- Shuai et al.[11], Wei and Hung[12], Daniel et al.[13] .Further approaches in Bing et al.[14] suggested dynamically learning CRFs for aspect extraction. Gaming datasets don't have any labeled aspects and it is hard to solve the problem using these approaches.

Moving over to unsupervised approaches, several variants of LDA models have been primarily used for unsupervised aspect extraction- Brody and Elhadad[15], Yan et al.[16],Mukherjee and Liu[17]. Another variant of unsupervised aspect extraction is generating co-occurring word pairs using the biterm topic model[18]. In [1], the unsupervised neural attention model has performed better than LDA, biterm model and k-means . In [19], Contrastive Attention is a single-head attention using an RBF kernel, word embeddings and POS tagger.

## 3 Approach

Figure 1 illustrates the overall workflow for the project. The project is divided into two parts:
1) Helpful review classification and 2) Game aspect extraction

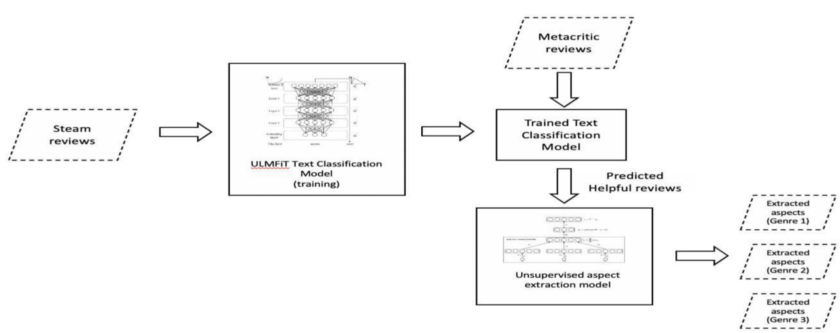


Figure 1: Overall Workflow Diagram

### 3.1 Game Review Classification

ULMFiT[2] model is used for predicting if a review is helpful or not. There are 3 steps to this modeling task:
1) LM Pre-training : The model is pre-trained on Wikitext 103 dataset which is a huge corpus of 103 Million words to learn language properties.
2) LM Fine-tuning: This is fine-tuning to the game reviews dataset since the words distribution is different. Discriminative fine-tuning and slanted triangular learning rates is incorporated to fine-tune the model.
3) Classifier Fine-tuning: This is done using gradual unfreezing of layers. The last layer is unfreezed and trained and then subsequent layers are unfreezed.

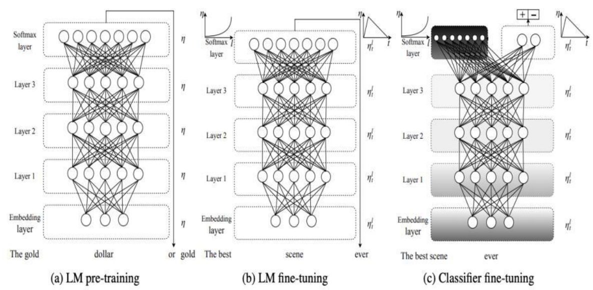


Figure 2: Unsupervised Neural Attention model Architecture[2]

### 3.2 Aspect Extraction of Game Reviews

#### 3.2.1 Architecture

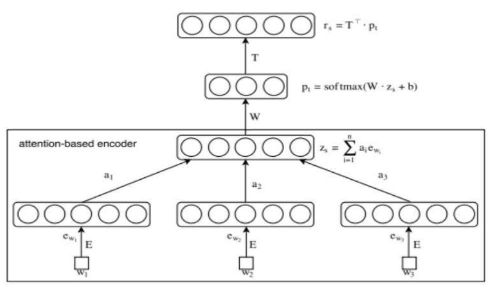


Figure 3: Unsupervised Neural Attention model Architecture

From Figure 1, each sentence in a review is encoded ($Z_s$) by using word embeddings of each word in the sentence as well as an attention weight which prioritizes the important words in a sentence.

$$z_s = \sum_{i=1}^{n} a_i e_{w_i}$$

$$a_i = \frac{exp(d_i)}{\sum_{j=1}^{n} exp(d_j)}$$

$$d_i = e_{w_i}^T \cdot M \cdot y_s$$

$$y_s = \frac{1}{n} \sum_{i=1}^{n} e_{w_i}$$

Given the sentence embeddings, the model works like an autoencoder, perform dimensionality reduction of the sentence embeddings from d dimensions to K dimensions and apply non-linearity function softmax to it. Further, the sentence embeddings are reconstructed using the aspect embeddings matrix and the dimensionality reduced sentence embeddings.

$$p_t = softmax(W \cdot z_s + b)$$

$$r_s = T^T \cdot p_t$$

#### 3.2.2 Loss function and Training

The objective is to minimize reconstruction error using contrastive max-margin objective function.

For training purposes, negative random samples of m sentences are picked from the training data and the word embeddings are averaged to form $n_i$. The goal is to maximize the inner product between $r_s$ and $z_s$ since $r_s$ is a reconstruction of $z_s$. It also minimizes the inner product between $r_s$ and $n_i$ since $r_s$ should not be similar to the randomly sampled sentence embeddings.

$$J(\theta) = \sum_{s \in D} \sum_{i=1}^{M} max(0, 1 - r_s z_s + r_s n_i)$$

Regularization is added to this to increase diversity of the aspect words.

## 4 Experiments

### 4.1 Game Review Classification

#### 4.1.1 Data

Steam game reviews[3] are used for the binary classification task of identifying if a review is helpful or not. The dataset has the game title,date of review,number of users that voted if a review is funny,number of users that voted if a review is helpful,number of hours played by the user and the review The dataset consists of 434,891 reviews. If a review has been upvoted by more than 1 user, the review is marked as helpful.

#### 4.1.2 Experimental details

The data split for training/dev/test set is 60:20:20. The stop words are removed and the text is converted to lowercase. The data is balanced by sampling the majority class.
For the LM fine-tuning model, gradually increase the learning rate and plot the loss to find the optimal learning rate. In this experiment, we found the optimal learning rate to be 1e-01. For the classifier fine-tuning, the last two layers were unfreezed and is trained for 4 epochs and then all the layers are unfreezed to train completely using differential learning rates.

#### 4.1.3 Results

The evaluation metric used is Accuracy since there is no class imbalance. The baseline model is a Naive Bayes model using unigrams. ULMFiT model performs significantly better than the other models.

| Model | Validation Accuracy |
|---|---|
| Naive Bayes-Unigrams | 62.7% |
| SVM -Unigrams | 60.9% |
| Logistic Regression-Unigrams | 62.2% |
| Naive Bayes-Bigrams | 63.3% |
| ULMFit | **91.8**% |

Table 1: Text Classification results

### 4.2 Aspect Extraction of Game Reviews

#### 4.2.1 Data

Metacritic game reviews[4] are used for aspect extraction task. The dataset has the game title, the platform it is available on, a user score, review by the user and the username. The dataset consists of 283,983 reviews.

The best model from Section 4.1.3 is used for predicting if the review by the user is helpful or not. The best model from the previous section predicted that 114,792 reviews are useful(40.4%).

For this task, we have divided the dataset by genre:

| Name | Description | Dataset size |
|---|---|---|
| All Genres | Reviews containing all genres | 114,792 |
| Action | Reviews containing only Action genre | 26,282 |
| Fantasy | Reviews containing only Fantasy genre | 11,154 |
| Sports | Reviews containing only Sports genre | 4,667 |

Table 2: Training datasets for aspect extraction

#### 4.2.2 Evaluation method

For aspect quality evaluation, a coherence score is used:

$$C(z; S^z) = \sum_{n=2}^{N} \sum_{l=1}^{n-1} log \frac{D_2(w_n^z, w_l^z) + 1}{D_1(w_l^z)}$$

Given an aspect z and a set of top N words of z, $S^z = w_1^z, ..., w_N^z$, the coherence score is calculated as follows: where D1(w) is the document frequency of word w and D2(w1, w2) is the co-document frequency of words w1 and w2. Higher the coherence score, the aspect words are more coherent.

#### 4.2.3 Experimental details

The experimenet setup is very similar to [1]. The word embedding matrix E is trained using word2vec with negative sampling on each dataset. The embedding size is set to 200 and negative sample size to 5. The aspect embedding matrix T is initialized with the centroids obtained from running k-means on word embeddings. During the training, the word embedding matrix E and is optimized using Adam with batch size of 50, learning rate 0.001 and for 15 epochs . We set the number of negative samples per input sample m to 20. The number of aspects is set at 10. The model has been experimented with varying aspect sizes of 10,15,20,30.

#### 4.2.4 Results

The baseline model is a LDA model. From Table 3, it can be seen that LDA consistently performs much worse than the neural attention model.
For the all genres dataset, k=20 has the highest coherence score.

| Dataset | Model | Number of aspects | Coherence score |
|---|---|---|---|
| All Genres | LDA | 10 | -5.0462 |
| All Genres | LDA | 10 | -5.2966 |
| All Genres | LDA | 20 | -5.5625 |
| All Genres | LDA | 30 | -7.4413 |
| All Genres | Neural Attention | 10 | -5.0112 |
| All Genres | Neural Attention | 15 | -4.0821 |
| All Genres | Neural Attention | 20 | **-3.6896** |
| All Genres | Neural Attention | 30 | -3.9196 |

Table 3: Aspect Extract experiment results for all genres dataset

For the Action, Fantasy and Sports dataset, k=10 has the highest coherence score.

| Dataset | Model | Number of aspects | Coherence score |
|---|---|---|---|
| Action | Neural Attention | 10 | **-3.3021** |
| Action | Neural Attention | 15 | -3.6309 |
| Action | Neural Attention | 20 | -4.6449 |
| Action | Neural Attention | 30 | -5.3422 |

Table 4: Aspect Extract experiment results for Action dataset

| Dataset | Model | Number of aspects | Coherence score |
|---|---|---|---|
| Sports | Neural Attention | 10 | **-1.853** |
| Sports | Neural Attention | 15 | -2.2168 |
| Sports | Neural Attention | 20 | -2.234 |
| Sports | Neural Attention | 30 | -3.0872 |

Table 5: Aspect Extract experiment results for Sports dataset

| Dataset | Model | Number of aspects | Coherence score |
|---|---|---|---|
| Fantasy | Neural Attention | 10 | **-4.5334** |
| Fantasy | Neural Attention | 15 | -4.9968 |
| Fantasy | Neural Attention | 20 | -5.7932 |
| Fantasy | Neural Attention | 30 | 6.8962 |

Table 6: Aspect Extract experiment results for Fantasy dataset

# 5 Analysis

## 5.1 Game Review Classification Analysis

Naive Bayes model predicts the below review as not helpful which is not ideal since this review has a lot of key information like cinematic effects,storyline and graphics. But, the ULMFiT model predicts this sentence as helpful. This is because the Naive Bayes model uses the game review corpus for prediction, whereas ULMFiT is a pretrained model. Cinematic effects and storyline did not exist in the Steam reviews corpus.

**Review:**
*"I don‘t know why but I have seen some people giving negative scores for this game. If they were wrong? YES, completely wrong! All I have to say is those people who gives low grades to this big of a game is either completely crazy, or is a hater, or do not know how to play the game and because of that gives it a low score. This game despite how old it is, still makes me fascinated with itsI don‘t*

*know why but I have seen some people giving negative scores for this game. If they were wrong? YES, completely wrong! All I have to say is those people who gives low grades to this big of a game is either completely crazy, or is a hater, or do not know how to play the game and because of that gives it a low score. This game despite how old it is, still makes me fascinated with its cinematic effects and graphics(by the way, graphics doesn't matter, what matters is the fun). I did not accomplished the entire game a.k.a. side quests, but I accomplished the main storyline and that was REALLY worth of my time, I simply loved this game, it made me a fan of the Zelda games, I highly recommend that you play it, you surely won't regret it, it's as the critics says a Masterpiece. And I have to say thanks to Shigeru Myamoto and Nintendo. Now if you excuse me, I have a new adventure to set on Zelda: Skyward Sword!"*

### 5.2 Aspect Extraction Analysis

From Table 7, it can be inferred that aspects extracted from the all genres dataset are very generic aspects and are not specific to any game. While some aspects like storyline, plot are pretty useful; aspects like people are not very useful for gamers.

| Inferred aspects | Aspects |
|---|---|
| Time | day, night, time, hour, year |
| Game version | franchise, genre, series, version, rpg |
| Game purchase behavior | gamer, fan, multiplayer, play, bought |
| Battle | dodge, enemy, attack, combo, trigger |
| Type of game | dlc, explasion, currency, addon, free |
| Gaming world detail | landscape, forest, scenery, art, detail |
| Story | story, plot, narrative, storyline, dialogue |
| Dislike | bad, horrible, terrible, suck, awful |
| Game names | halo, warcraft, cod, mario, zelda |
| People | customer, company, community, fans, forums |

Table 7: Aspects extracted for All Genres dataset

To improve the analysis, the dataset had been split by genre. It can be seen that, the aspects like enemy,gameplay world, actions are more relevant for end users.

| Inferred aspects | Aspects |
|---|---|
| Protagonist | protagonist, character, villain, hero, historia |
| Game Characters | drake, ellie, nathan, zelda, hyrule |
| Exploration | world, explore, journey, adventure, wild |
| Storyline | story, narrative, plot, storytelling, twist |
| Actions | climbing, bow, crafting, walk, jump |
| Degree of dislike | repetitive, boring, tedious, pointless, dull |
| Game quests | mission, activity, quest, objective, variety |
| Enemy | enemy, attack, gun, kill, skill |
| Video game components | texture, animation, graphic, lighting, music |
| Gameplay world | city, wall, building, mountain, car |

Table 8: Aspects extracted for Action dataset

For Sports, lesser number of aspects were interpretable, but listing some that could be useful for end users:

| Inferred aspects | Aspects |
|---|---|
| Object | ball, football, basketball |
| Franchises | NBA, Madden, FIFA, NHL |
| Sports actions | pass, shot, tackle, run, move, goal |

Table 9: Aspects extracted for Sports dataset

But, for fantasy and sports, we could infer very few aspects since the dataset size was very small. Also, since, we don't want to remove any gaming terms, non-English words were not removed. For example, if the English words are removed, game names like Zelda and Mario would be removed. Due to this, a lot of non-English words were grouped together as aspects which is not good:

| Inferred aspects | Aspects |
|---|---|
| Russian | не, на, что, игра, это, но, из, очень, по, как |
| Spanish | una, que, para, por, pero, su, la, lo, jogo, en |

Table 10: Incorrect aspects

## 6 Conclusion

ULMFiT model has been implemented for identifying helpful reviews. The unsupervised neural attention model has been implemented successfully for game review aspect extraction. Further, splitting the dataset by genre has helped in building more interpretable aspects. The limitations of this approach is that the dataset consists of multiple languages and at times, words of the same language are getting grouped together. But, considering only English words would not be helpful for the gaming dataset. For future work, we need to identify certain gaming words that should not be omitted and consider only English words.

## References


[1] Hwee Tou Ng Daniel Dahlmeier Ruidan He, Wee Sun Lee. An unsupervised neural attention model for aspect extraction. In *Association for Computational Linguistics*, 2017.

[2] Sebastian Ruder Jeremy Howard. Universal language model fine-tuning for text classification. In *Association for Computational Linguistics*, 2018.

[3] https://www.kaggle.com/luthfim/steam-reviews-dataset.

[4] https://www.kaggle.com/dahlia25/metacritic-video-game-comments.

[5] https://github.com/fastai.

[6] https://github.com/ruidan/unsupervised-aspect-extraction.

[7] https://github.com/kirillkrasikov/attention-and-capsule-based-aspect-extraction.

[8] https://www.marketwatch.com/story/videogames-are-a-bigger-industry-than-sports-and-movies-combined-thanks-to-the-pandemic-11608654990.

[9] https://en.wikipedia.org/wiki/list_of_playstation_games_(a%e2%80%93l).

[10] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.

[11] Shuai Wang Lei Zhang and Bing Liu. Deep learning for sentiment analysis: A survey. 2018.

[12] Wei Jin and Hung Hay Ho. A novel lexicalized hmm-based learning framework for web opinion mining. In *In Proceedings of the 26th International Conference on Machine Learning*, 2009.

[13] Daniel Dahlmeier Wenya Wang, Sinno J. Pan and Xiaokui Xiao. Aspect level sentiment classification with deep memory network. In *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

[14] Bing Liu Lei Shu, Hu Xu. Lifelong learning crf for supervised aspect extraction. In *Association for Computational Linguistics*, 2017.

[15] Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Association for Computational Linguistics*, 2010.

[16] Hongfei Yan Wayne Xin Zhao, Jing Jiang and Xiaoming Li. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.

[17] Arjun Mukherjee and Bing Liu. Aspect extraction through semi-supervised modeling. In *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistic*, 2012.

[18] Yanyan Lan Xiaohui Yan, Jiafeng Guo and Xueqi Cheng. A biterm topic model for short texts. In *In Proceedings of the 22nd International World Wide Web Conference.*, 2013.

[19] Andreas van Cranenburgh Stéphan Tulkens. Embarrassingly simple unsupervised aspect extraction. In *Association for Computational Linguistics*, 2020.

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19]