Using User Created Game Reviews for Sentiment Analysis: A Method for Researching User Attitudes

Björn Strååt
Stockholm University
Department of Computer and Systems Sciences
Stockholm, Sweden
bjor-str@dsv.su.se

Harko Verhagen
Stockholm University
Department of Computer and Systems Sciences
Stockholm, Sweden
harko.verhagen@dsv.su.se

ABSTRACT
This paper presents a method for gathering and evaluating user attitudes towards previously released video games. All user reviews from two video game franchise were collected. The most frequently mentioned words of the games were derived from this dataset through word frequency analysis. The words, called "aspects" were then further analyzed through a manual aspect based sentiment analysis. The final analysis show that the rating of user review to a high degree correlate with the sentiment of the aspect in question, if the data set is large enough. This knowledge is valuable for a developer who wishes to learn more about previous games success or failure factors.

Author Keywords
Sentiment; sentiment analysis; user created content; reviews

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Evaluation/methodology

INTRODUCTION
It is commonly acknowledged that designers and developers have much to gain from knowing the needs and expectations of their future customers and users. In interaction design, many methods of exploring this exist; interviews, observations, surveys, and other techniques exist both in the industry and the academic world [1].

In relatively recent times, customer/user attitude researchers have turned to user created content, such as social media, internet forum, and user reviews, with the intent of mining user attitudes from within. Online text content is not a new source, but the phenomenon was earlier more focused on expert rather than user created content. Users often express themselves regarding their experiences; many services provide user ratings and review services, e.g. products on amazon.com online store, tourist guides such as Yelp.com, and TripAdvisor.com, movie reviews such as rottentomatoes.com, and many more. The video game community is no different. The content provider service Steam allows the users to vote and comment on games, and the website Metacritic.com present both expert- and user created reviews. User created content offers a vast and varied source of data for anyone who wish to explore the user sentiment beyond the basic rating of previously released products.

In this study, we have performed an Aspect Based Sentiment Analysis (ABSA) [2] based on data gathered from user reviews regarding two video game series, on Metacritic.com. Our purpose was to explore if the sentiment an aspect (commonly used words in the reviews) was used in, would reflect the overall rating from the reviewers. A positive result would imply that user reviews can be used to explain user attitudes (positive/negative sentiment) from a root-cause point of view (the aspects).

The results show that, given that the data set is extensive enough, there is a strong connection between the sentiment of the aspect and the rating the reviewer provided.

BACKGROUND
The use of video game reviews as a resource for game studies is not a common phenomenon. Most of the studies that has been performed, has been on professional reviews: Pinelle, Wong & Stach [3] used professional reviews as a source to find common video game issues, which they compiled into a set of design patterns, Zagal, Ladd & Johnson [4] found that game reviews often include design suggestions and serious discussions on game designer's intention and goals. User created reviews has been used as well, but not as frequently: Strååt & Verhagen [5] used user reviews to evaluate video game heuristics, Zagal & Tomuro [6] studied cultural differences and similarities in user created reviews from Japan and USA, and quite recently, Koehler, Arnold, Greenhalgh, Owens Boltz & Burdell's published their article "*A Taxonomy Approach to Studying How Gamers Review Games*" [7]. They used an existing theoretical model, a video game taxonomy, and compared user submitted reviews with the categories of the taxonomy. They found that users to a certain degree used the same

GHITALY17: 1st Workshop on Games-Human Interaction, April 18th, 2017, Cagliari, Italy.

concepts as the taxonomy, and that there was a difference in use of the concepts depending on the game rating. As more researchers move into the field, we would like to propose our method as presented in this paper.

Metacritic

Metacritic.com is a site that aggregates professional reviewer scores from various online media review sources. Television shows, movies, music and video games (various platforms) are examples of media that are presented. Metacritic calculates an average score called Metascore, based on the various professional reviewers by converting the reviewers' local score into a score of 0 to 100 (e.g. a local score of 8 out of 10 renders a Metascore of 80). These scores are weighted (based on the quality and overall stature of the source) and finalized into a professional Metascore.

Regular non-professional users are also allowed to score the media on a scale of 0 to 10. The unweighted average of this score is presented by Metacritic as the Userscore. Non-professional users can also post their own reviews along with their score. The User score does not consider the length or quality of these reviews; a simple four-word comment, such as "this game is good", is valued the same as an analytical 500-word essay. User reviews and scores are posted anonymously under a self-selected user name. The user score is divided into three tiers: Positive, Neutral and Negative, where Positive is ratings 8 to 10, Neutral is ratings 5 to 7, and Negative is ratings 0 to 4. The rating tiers are color coded in green for Positive, yellow for Neutral and red for Negative.

Metacritic has been the subject of many discussions. The validity and value of the professional reviews have been questioned in various video game blogs and online magazines [8] [9], and the site has been used in game and social studies, e.g. as an examination and comparison of player experience vis-à-vis professional reviews [10], or as a key factor in assessing game value and quality [11]. Most commonly, the discussion has been around the professional reviews. In this study however, we have only looked at the User score and user comments.

Games in this study

The goal of this study is to see if the user sentiment differs between games that are released in a series. To this end, we decided to examine the user comments of the game series "Dragon Age" and "Mass Effect". At the time of the study, Dragon Age has three installments: Dragon Age: Origin (DA1) [12], Dragon Age 2 (DA2) [13], and Dragon Age: Inquisition (DA3) [14]. Mass Effect has four installments, but only the three first existed when we performed the data collection. These are Mass Effect (ME1) [15], Mass Effect 2 (ME2) [16], Mass Effect 3 (ME3) [17].

We chose these franchises since they are widely known, and represents a relatively common and popular game genre (role playing games), and most importantly, they have received varying ratings from players. The PC version of DA1 received 8.7/10.0 userscore on Metacritic, DA2 received 4.5/10.0, and DA3 received 5.9/10.0.

ME1 received a userscore of 8.6/10.0, ME2 received 8.8/10.0, and ME3 was rated 5.6/10.0.

The sudden drop in ratings from DA1 to DA2, and ME1/2 to ME3 tells us that something has changed in the series, either with the games or the users. This is the phenomenon we wanted to explore by analyzing the user reviews.

METHOD

In this section, we describe our scientific approach and methods for data gathering and analysis. We use a qualitatively driven mixed methods approach, where quantitative methods supplement and improve the study's results. The qualitative analysis is done through a through manual aspect based sentiment analysis. The quantitative analysis was done through hypothesis testing using a Chi-square test.

Aspect Based Sentiment Analysis

An aspect based sentiment analysis (ABSA) [2] is performed when user sentiment of certain aspects of a multi-aspect entity is to be measured, in a dataset gathered from user comments, such as online forum or user created reviews. Video games have plenty of aspects that the user considers when playing, e.g. playability, graphics, storyline.

Aspects are words or phrases that exist either explicit or implicit in the dataset. Explicit aspects are the actual word in context, and implicit aspects are inferred from the context. For example, if the aspect is *gameplay*, an explicit occurrence could be "*I really enjoyed the gameplay*", and an intrinsic could be "*I really enjoyed the challenges and the features of X*."

The aspects are determined through a word frequency analysis. After the dataset is collected, product or domain relevant words that occur on a frequency above a pre-set threshold are retained for the following sentiment analysis step. The sentiment analysis is then performed either through a scripted natural language processing algorithm, or through a manual read through. The result will show the sentiment for each aspect, for example in terms of positive, neutral, or negative sentiment.

Word frequency and selection

The data collection for our ABSA was performed in the following steps. First, we collected all user reviews on the PC-version of the three games from the Dragon Age franchise: DA1, DA2, and DA3, and the three first games from the Mass Effect franchise: ME1, ME2, ME3, from Metacritic.com. As mentioned in the Metacritic description in the background section, Metacritic authors rate their own reviews to reflect their experience of the game in question. This is a rating from 0 to 10, but in effect it will categorize the comment as one of three tiers: low, medium, or high rated. We decided to only work with the reviews of the PC-

version (the games exist for multiple platforms) as it was the versions that we were familiar with.

For each game, we did a word frequency analysis, using AntConc [1], to find which aspect that was most frequently used in the reviews. As we had no previous practice of this method in this context, the threshold was set after we saw the results – we decided to pursue the three most frequent explicit aspects that were shared by all three games. These explicit aspects were: *Story*, *Combat*, and *Character*. All reviews that did not contain any of the aspects were omitted from the dataset. As the reviews were rated by the authors, we already had the rating categories.

Since the review rating and the sentiment of the aspect may differ – for example, a high rating review may use an aspect in a negative way – it was important to collect all reviews of all ratings, that contained at least one aspect. *Figure 1* is an illustration on how frequent the aspects were in relation to review rating. As can be seen, the aspects tend to be more frequent in low rated reviews than high and mid rated reviews. This was true for all games, but for reasons of limited space, only one figure is shown.

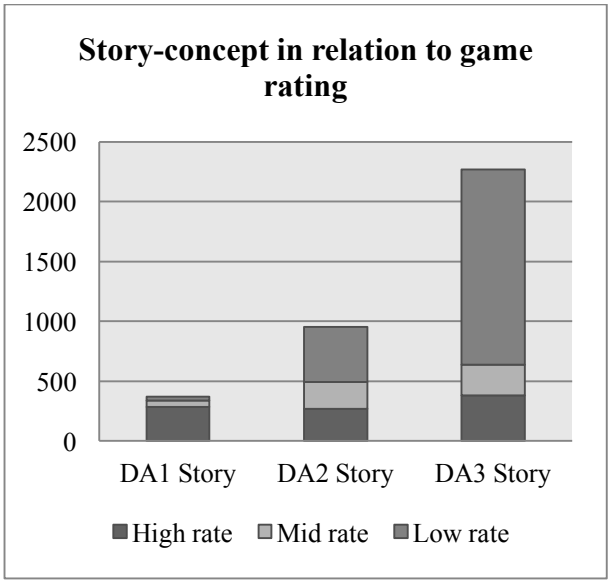

Figure 1: Relation between aspect and review rating

After the data collection, we had a dataset of reviews for each game, regarding the three aspects (story, combat, character). Each review was categorized into its original rating level.

So, in conclusion of this section:

- Aspects were determined through word frequency analysis of all the user reviews
- The three most frequent aspects were combat, story, character.
- Each game had a number of reviews
- A review contains at least one of the aspects
- A review is rated as either low, medium, or high
- The dataset contains all reviews, sorted by game, rating, and aspect.

Manual Sentiment Analysis

The sentiment analysis was performed online, through an online crowdsourcing service.[2] The rating and name of the game was omitted for the evaluators to limit the risk of bias. The evaluators were asked to read a review, or excerpt of a review, which contained one of the aspects, and to determine if the author of the review had used the aspect in a positive, neutral, or negative way. The following quote is an example of an excerpt that the evaluators judged:

"The menus, crafting and combat are so totally and completely cumbersome. Everything is very statically organized and takes so much time. I spent an ungodly amount of hours collecting resources, crafting things, comparing items to what I already owned and it is just so, so, so cumbersome and tiresome, it really damages the game"

The aspect of combat occurs in the quote, and the overall use of the aspect is considered negative.8268 review excerpts from the DA series and 3357 from the ME series were analyzed this way, and each aspect was judged by at least three evaluators. If an excerpt would contain more than one aspect, it would be run again, through a second (or third) sentiment analysis, where that aspect would be in focus for the evaluator. When the sentiment analysis was done, the dataset was reconstructed with rating and game name.

Chi-square analysis

Chi square is a common test for hypothesis testing. At its core, it calculates the differences between observed frequencies and expected frequencies in a row by row and column by column calculation, adding the calculations for each cell together into one comprehensive measure. Depending on the degrees of freedom (number of rows minus 1 times number of columns minus one) and the measure of reliability, cut-off measures have been calculated. A Chi square above the cut-off value means that the probability of the variables to be independent (Null hypothesis) is below the reliability (usually .05 or lower). In general, for 2*2 tables, a lower threshold of 5 for each expected frequency is thought to be needed, even if some

[1] AntConc, by Anthony (2012), is a freeware concordance and text analysis tool by Dr Laurence Anthony at the Faculty of Science and Engineering at Waseda University, Japan (http://www.antlab.sci.waseda.ac.jp/index.html).

[2] www.crowdflower.com; a data mining and crowdsourcing service where researchers can upload their data e.g. for manual sentiment analysis by anonymous evaluators.

debate exists concerning this value. Thus, for a 3*3 table such as ours, at least 45 observations need to exist from the start for Chi square to be a reliable test by general agreement.

RESULT

After the sentiment analysis, we processed the data from an analytical standpoint. Table 1 shows the complete data set for all three DA games, distributed on review ratings, aspects and sentiment, and table 2 shows the same for the ME games.

We tested the relevance of each of the three aspects for the overall review. We constructed the following null hypothesis: *There is no relationship between the values of aspect X (character, combat or story) and the overall review rating.*

Dragon Age		**Review rating**				
Aspect		**Low**	**Mid**	**High**		
Char.	bad	633	257	87	977	
	neutral	1038	72	92	1202	
	good	68	148	543	759	2938
Comb.	bad	520	211	72	803	
	neutral	358	50	48	456	
	good	43	83	353	479	1738
Story	bad	993	278	69	1340	
	neutral	1056	119	129	1304	
	good	72	142	734	948	3592
		4781	1360	2127	8268	

Table 1: The aspects distributed on review ratings, for all three games in the Dragon Age franchise. The values are from the evaluators sentiment analysis.

Mass Effect		**Review rating**				
Aspect		**Low**	**Mid**	**High**		
Char.	bad	256	120	33	409	
	neutral	28	34	49	111	
	good	53	77	411	541	1061
Comb.	bad	88	66	35	189	
	neutral	10	19	80	109	
	good	25	45	191	261	559
Story	bad	425	164	65	654	
	neutral	59	70	128	257	
	good	58	91	667	826	1737
		1012	686	1659	3357	

Table 2: The aspects distributed on review ratings, for all three games in the Mass Effect franchise. The values are from the evaluators sentiment analysis.

Chi Square Test Results

Using the Chi square test, we obtained the values presented in table 3 and 4. The tables show Chi-square per aspect for each game.

	Aspect	**Chi square**
DA1	Character	120,2
	Combat	100,4
	Story	196,6
DA2	Character	304,9
	Combat	299,6
	Story	426,6
DA3	Character	1072,5
	Combat	374,4
	Story	1250,2
DA series	Character	1541,3
	Combat	813,8
	Story	1963,4

Table 3: Chi-square values for each aspect from the DA series

All values exceed the threshold at p= 0.001 and 4 degrees of freedom (18,465) thus in all cases of the DA series; the null hypothesis can be reject. We conclude that there is a correlation between the aspect value and the overall review value.

	Aspect	**Chi square**
ME1	Character	94.12
	Combat	31.15
	Story	139.48
ME2	Character	252.13
	Combat	75.20
	Story	470.90
ME3	Character	466.21
	Combat	163.41
	Story	797.27

Table 4: Chi-square values for each aspect from the ME series

Given the minimum value of 5, each row or column should have at least 15 observations, which in the case of ME1 does not hold for any of the aspects as there are fewer than 15 observations in the "Low" column for each of the 3 aspects. The same goes for ME2 for the Combat aspect (only 5 with score "Low" and only 7 with score "Mid").

DISCUSSION

Our results show that if an aspect occurs in a review, the sentiment of that aspect will reflect the rating of the review. The null hypothesis was falsified for all games, and all aspects except for two of the games in the ME series, ME1 and ME2.

This implies that the aspects reflect areas, in the games, that are disliked by the users. The relatively high frequency of the aspects is an indication that these areas are the most important ones for the users. It also indicates that the root cause of the low rated reviews is to be found within the game features that the aspects represent.

The null hypothesis was not possible to falsify for ME1 and ME2 due to the lack of data for these two games. Looking at table 2, we can see that it only exists 10 Low review/Combat neutral, meaning that this data point cannot be calculated using Chi-square. This is a good indication that the threshold for the word frequency analysis (please see method section, word frequency and selection) must be at least 45 for the analysis to be valid.

However, a game designer might not need the analysis to be statistically valid: Consider figure 1. The amount of user reviews increase for each instalment of the game franchise, but a large majority of the increase is within the negatively rated reviews. This is our first clue that the related aspect is important to the users. This is not a statistically validated result, but it gives us an indication if we are looking at something that needs to be further investigated. The amount of low rated reviews that contain at least one of each aspect may indicate that these aspects are part of the reasons that users didn't appreciate the games. From a video game developer standpoint, we could stop here. It wouldn't take too long to manually read through a few pages of these comments to get an estimated overview whether the aspects are used in a negative sentiment or not. A developer can, at this stage, get this overview and regard their design choices accordingly.

The frequency of the aspects implies that they are important to the users – this implies that the authors of the low rated reviews are disappointed of the aspects as presented in the games. A future research task would be to perform a more qualitative analysis, on user review level, to pinpoint the root cause of the problems that the users experience. A content analysis, for example, of the material would give a more detailed insight. Furthermore, we have only worked with the PC-reviews of the game franchises. A full out analysis of all the platforms for all the games would possibly render a different result, or enhance the one presented in this paper.

REFERENCES

[1] D. Benyon, P. Turner and S. Turner, Designing interactive systems: People, activities, contexts, technologies, Pearson Education, 2005.

[2] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos and S. Manandhar, "Semeval-2014 task 4: Aspect based sentiment analysis," *Proceedings of SemEval,* pp. 27-35, 2014.

[3] D. Pinelle, N. Wong and T. Stach, "Heuristic Evaluation for Games: Usability Principles for Video Game Design," in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2008)*, 2008.

[4] J. P. Zagal, A. Ladd and T. Johnson, "Characterizing and understanding game reviews," in *Proceedings of the 4th international Conference on Foundations of Digital Games*, 2009.

[5] B. Stråat and H. Verhagen, "VOX POPULI - A Case Study of User Comments on Contemporary Video Games in Relation to Video Game Heuristics?," United Kingdoms, 2014.

[6] J. P. Zagal and N. Tomuro, "Cultural differences in game appreciation: A study of player game reviews," in *FDG*, 2013.

[7] M. J. Koehler, B. Arnold, S. P. Greenhalgh, L. O. Boltz and G. P. Burdell, "A taxonomy approach to studying how gamers review games," *Simulation & Gaming,* vol. 48, no. 3, pp. 363--380, 2017.

[8] "Metacritic Matters: How Review Scores Hurt Video Games," 08 08 2015. [Online]. Available: http://kotaku.com/metacritic-matters-how-review-scores-hurt-video-games-472462218. [Accessed 18 04 2016].

[9] "Time to kill Metacritic," 15 10 2014. [Online]. Available: http://www.mcvuk.com/news/read/time-to-kill-metacritic/0139824. [Accessed 18 04 2016].

[10] D. Johnson, C. Watling, J. Gardner and L. Nacke, "The edge of glory: The relationship between metacritic scores and player experience.," in *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play*, 2014.

[11] A. Greenwood-Ericksen, S. R. Poorman and R. Papp, "On the Validity of Metacritic in Assessing Game Value," *Eludamos. Journal for computer Game*

Culture, vol. 7, no. 1, pp. 101-127, 2013.

[12] *Dragon Age:Origins,* BioWare, 2009.

[13] *Dragon Age II,* BioWare, 2011.

[14] *Dragon Age: Inquisition,* BioWare, 2014.

[15] BioWare, *Mass Effect,* USA: Electronic Arts, 2007.

[16] BioWare, *Mass Effect 2,* USA: Electronic Arts, 2010.

[17] BioWare, *Mass Effect 3,* USA: Electronic Arts, 2012.