The Wisdom of the Gaming Crowd

ROBERT JEFFREY, PENGZE BIAN, FAN JI, and PENNY SWEETSER, Australian National University, Australia

In this paper, we report on three projects in which we are applying natural language processing techniques to analyse video game reviews. We present our process, techniques, and progress for extracting and analysing player reviews from the gaming platform Steam. Analysing video game reviews presents great opportunity to assist players to choose games to buy, to help developers to improve their games, and to aid researchers in further understanding player experience in video games. With limited previous research that specifically focuses on game reviews, we aim to provide a baseline for future research to tackle some of the key challenges. Our work shows promise for using natural language processing techniques to automatically identify features, sentiment, and spam in video game reviews on the Steam platform.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; • **Software and its engineering** → **Interactive games**; • **Applied computing** → **Computer games**.

Additional Key Words and Phrases: game reviews; natural language processing; NLP; video games; qualitative analysis; text mining; Steam

ACM Reference Format:

Robert Jeffrey, Pengze Bian, Fan Ji, and Penny Sweetser. 2020. The Wisdom of the Gaming Crowd. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '20 EA)*, November 2–4, 2020, Virtual Event, Canada. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3383668.3419915>

1 INTRODUCTION

Using natural language processing (NLP) techniques to analyse video game reviews offers a wealth of opportunities and value for video game developers, researchers, and players. With the video game industry expected to exceed \$160 billion in revenue in 2020 [31], assisting customers in making purchasing decisions is a valuable research avenue. In addition, helping the developers of video games understand the content of their product’s reviews could help them better address the feedback of their customers. Furthermore, we could glean insights from game reviews about how to design more enjoyable games in general, as has been attempted in previous research [4, 28, 29].

Video games, in comparison to consumer goods and services, have a few traits that make them unique. One key factor is the increased subjectivity of a consumer’s opinion of a video game. While customers have differing opinions about consumer products (e.g., electronics), these products often have objective features that make reviewing less subjective. In comparison, video games are a form of entertainment and personal taste plays a greater role. Additionally, video game reviews are often less formal and frequently contain jokes or humour, which makes it challenging for many NLP techniques.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

Table 1. A selection of metadata values provided by the Steam API.

Attribute	Description
review	Review text
voted_up	True if positive recommendation
votes_up	#users who marked as “helpful”
votes_funny	#users who marked as “funny”
comment_count	#comments on review

With over 30,000 games on gaming platform Steam, with some games having hundreds of thousands of reviews, it is not possible for players to digest all the available information before making an informed decision about what to buy. At the same time, Steam presents an opportunity to access large datasets of reviews for analysis and testing the application of NLP techniques. In this paper, we present the preliminary results of three projects in which we have applied NLP techniques to Steam reviews to attempt to (1) summarise game features, (2) determine positive or negative sentiment, and (3) identify spam reviews.

There has been limited previous research that specifically applies NLP techniques to the analysis of game reviews. Previous research has included using word frequency to analyse consumer attitudes towards video games [27], attempting to predict whether a game will be successful from its reviews [30], and comparing features of game reviews to other types of reviews [12].

The aim of the work presented in this paper is to tackle a set of challenges related to automatic analysis of video game reviews and to provide a starting point for future research in this area. We aim to provide tools, processes, datasets, and avenues for future research, so that we, as a field of games researchers, can make use of the extensive data contained in Steam (and other game) reviews to extract, analyse, and understand the wisdom of the gaming crowd.

2 STEAM REVIEWS

Steam is a distribution platform for video games and related software, developed by Valve Corporation. The platform has a strong community, including substantial participation in the reviewing of games. Reviews on the site are composed of two main elements: a binary “recommended?” variable, and the review text. Other users are able to rate reviews as either “helpful” or “unhelpful”, as well as to tag reviews as “funny”. It is not uncommon for applications to have thousands or tens of thousands of reviews, with some having over one million reviews [26]. These features, in combination with its powerful API, makes Steam ideal for our purposes. Previous research that makes use of the Steam platform has involved predicting sentiment of Steam reviews based on review text [32] and classifying reviews as “helpful” or “unhelpful” using the user-assigned “helpfulness” score as ground truth [18].

We used the official Steam API in order to collect our review datasets. The Steam API provides the review text alongside metadata about the reviews. A selection of useful metadata values are provided in Table 1, with the full list available in the Steamworks documentation [11]. Through these Steam API calls, we were able to collect approximately 150 reviews per second.

3 FEATURE SUMMARISATION

Automatic summarisation aims to condense a large corpus of text down into an easily human-readable summary. Often this summary takes the form of a paragraph of text, but in the case of opinion mining of reviews, it can also provide a

Table 2. Players and percentage of relevant features (%) for each game: Dota 2, PlayerUnknown’s Battlegrounds (PUBG), Counter-Strike: Global Offensive (CS:GO), Rainbow Six: Siege (R6:S), and Warframe.

Games	Players	% Features
Dota 2	1,033,925	53.3%
PUBG	931,407	60.0%
CS:GO	680,071	53.3%
R6:S	134,776	66.7%
Warframe	115,102	66.7%

detailed breakdown of people’s opinions on a product. Feature-based summarisation research has drawn reviews from large review sites including e-commerce platform Amazon [1, 14], product review website CNET [14], travel review website TripAdvisor [21], and business review site Yelp [10]. While research has been conducted in analysing video game reviews [7, 18, 32], there is limited research into summarising these reviews. The aim of this project is to perform a preliminary investigation of feature summarisation for video game reviews.

Feature summaries are composed of two main elements [14]. First, commonly discussed aspects of the product, called "features", are extracted from the review dataset. Next, each feature is given a "sentiment" composed of the percentage of reviews that discuss the feature which display positive sentiment. Our summarisation technique is based upon the algorithm described in [8], with some modifications. Our sentiment scoring algorithm works similarly to the High Adjective Count (HAC) algorithm, we find each adjective and the noun associated with it (i.e., the nearest noun). We then create a four word "context" composed of the two words before the adjective (if they exist), the adjective itself, and the nearest noun. This context is then scored using the VADER algorithm [15] as implemented in the NLTK library [3] as either positive or negative. Finally, we determine the features to use by selecting the features with the highest HAC scores. We can then present these features alongside the scores we find using our sentiment analysers.

We constructed a review corpus by scraping Steam reviews for the five games with the highest player count at the time of our experiment (see Table 2). In this preliminary study, we selected games with a large number of reviews, but future work should include a broad selection of games. In order to evaluate our algorithm, we generated feature summaries of the five games we selected, and qualitatively analysed the top 15 features of each summary. We qualitatively marked each feature as "relevant" or "irrelevant" and computed the percentage of reviews that were "relevant" (see Table 2). Relevance encompassed both relevance to the game itself, and the property of being specific enough to provide relevant information to a reader. For example, the title of each game was not marked as relevant.

Our results suggest that the existing algorithms are reasonably effective but have faults. The majority of features were useful, both having more generic features such as "gameplay" and "graphics" as well as features specific to the games themselves, such as "weapons" for Warframe. However, there was still a reasonable amount of non-useful terms, both with non-specific features such as titles or genres of the game, as well as difficult to interpret features such as "life" or "operators".

One major flaw was that the sentiment analyser was not trained on our dataset, and thus did not have full knowledge of the video game context. For example, the phrases "too many cheaters" and "too many hackers" have similar meanings in a video game context. However, the analyser rated the former as negative and the latter as neutral in our summary of Counter Strike: Global Offensive. Another issue is the susceptibility of the algorithm to spam due to the occurrence based HAC algorithm. For example, in PlayerUnknown’s Battlegrounds reviews, the phrase "#regionlockchina" was repeated in numerous reviews.

Table 3. Review count, positive (+) % reviews, and prediction accuracy of Naïve Bayes (NB) and Support Vector Machine (SVM) on reviews for each game: Just Cause 4 (JC4), Totally Reliable Delivery Service Beta (TRDS), Red Alert 3 (RA3), and No Man’s Sky (NMS).

Games	Reviews	+%	NB	SVM
JC4 2	4050	42%	83.8%	83.1%
TRDS	1004	90%	93.2%	93.4%
RA3	2167	80%	86.2%	87.9%
NMS	81979	53%	83.1%	80.2%

4 SENTIMENT ANALYSIS

Sentiment analysis aims to identify and extract subjective meaning in the target contextual material [5]. Sentiment analysis systems are widely implemented as emotions are fundamental to nearly all people related activities and are essential influencers of our behaviours. Our views and interpretations of reality, as well as our decisions, are largely dependent on how others see and judge the world, which is why we seek advice when we need to make a decision.

Our aim in this project is to apply NLP techniques to Steam reviews to attempt to predict whether the game review text is positive or negative. In this project, we applied two supervised machine learning techniques, Naïve Bayes and Support Vector Machine algorithms, to analyse Steam reviews. We used standard implementations of the algorithms in the Python machine learning module, scikit-learn. We analysed the performance of the algorithms on Steam reviews for four games. We selected games with at least 1000 reviews and a mix of games with mostly positive or mixed reviews.

The results (see Table 3) suggest that the algorithms are reasonably effective, but that there are still opportunities for further enhancements. We found that games with a higher percentage of positive reviews had higher prediction accuracy for both algorithms. Conversely, if reviews are closer to an even split of positive-negative, then it is more difficult to predict the sentiment.

The Naïve Bayes (NB) algorithm is simple but effective. Its computational complexity grows linearly as the number of samples or features grow, which makes the speed of training and prediction very fast. When the number of features is constant, even if there are more than 1,000,000 reviews in the database, NB still works well. However, NB assumes that features are independent and ignores inter-feature interactions. For example, “this game is not bad” shows a positive opinion, but NB independently processes the features “not” and “bad”, which make the review seem negative.

For the Support Vector Machine (SVM) algorithm, the computational complexity increases exponentially as the number of samples grows. This limits the SVM to small-scale data, but in a high-dimensional feature space. Compared to NB, by the way of converting a sample into a high-dimensional vector, SVM includes the interaction between features to some extent. However, when we encounter a situation where the data set is not linearly separable, we need to find a suitable kernel function to replace the linear kernel function. We need to keep trying different parameters and to find the best ones through continuous testing.

Based on our experiments, we can propose some suggestions that can improve the performance of the model:

- First, the interaction between tokens (e.g., words, punctuation) could be overlooked as each feature consists of only one token. When extracting features, we can select some features consisting of two adjacent tokens.
- Second, we should weight the influence of different words on sentiment analysis according to their parts-of-speech [2]. Adjectives and adverbs are often subjective and can directly indicate whether a review is positive or negative. Nouns tend to be more objective and indirectly express opinions.

- Third, when we tested different data sets, it is difficult to find the optimal number of extracted features. We usually need to try different numbers to approach the best value. The performance could also be improved by using a pre-trained model and constant parameters to extract features, such as the BERT model [6].

5 SPAM REVIEWS

The potential value of online reviews has led to more and more spam reviews appearing on the web. These spam reviews are widely distributed, harmful, and difficult to identify manually [24]. Websites for accommodation, travel, and shopping are plagued by spam reviews [19], accounting for about 14-20% of reviews on Yelp [9, 22]. Spam reviews fall into three categories [16]: deceptive or untruthful, reviews on brands, and non-reviews (advertisements, irrelevant reviews, and meaningless reviews).

In this project, we explore and implement generalised approaches for identifying online deceptive spam game reviews from Steam. Due to the similarity between game reviews and most text comments, we can use many mature NLP technologies and machine learning methods to analyse and classify game reviews. However, some unique characteristics of game reviews require further analysis, as there is a lack of research in this area.

This project analyses spam game reviews and presents some techniques to detect them. We attempt to validate the application of existing technology to the detection of spam game reviews. In addition, we aim to identify the unique features of game reviews and to create a labelled game review dataset based on different features. We made use of three review datasets: (1) 33,000 unlabelled Steam game reviews we collected from 8 different games, (2) the Gold-Standard deceptive opinion dataset comprised of 1,600 labelled reviews from various websites [23, 24], and (3) 100,000 Steam game reviews with sentiment labels [25].

We divided spam game reviews into three different categories: untruthful reviews, duplicate reviews, and non-reviews. Untruthful reviews, or deceptive reviews, are malicious and generally difficult to detect manually. Duplicate reviews have the same semantics, same polarity and similar length, but different users. Non-reviews refer to reviews that have no meaning or effect. We have identified the following unique features of game reviews, based on the Steam datasets:

- (1) Game reviews contains more symbols or emojis, as Steam allows players to add different emojis in reviews.
- (2) Steam does not limit the length of reviews, which means the distribution of length of reviews is scattered. Using the average length as a feature indicator is not very helpful, but longer reviews have more credibility.
- (3) The emotional polarity of game reviews is easy to judge. The rating for each review can be used for predicting the polarity.
- (4) Game reviews contain more game terms, irregular words, and network terms.

For the three kinds of spam game reviews we defined:

- the non-reviews were detected via pre-processing (e.g., no meaningful words, high ratio of symbols);
- the duplicate reviews were detected using the Doc2Vec algorithm [17] to represent the similarity of reviews by distance;
- for the untruthful reviews, we used a Positive Unlabelled Learning [13, 20] technique to train a decision tree classifier, which we applied to the Gold-Standard dataset to verify the validity prior to applying to the unlabelled Steam reviews. Our accuracy peaked at 72.2%, compared to human accuracy of around 60% [24].

Through this project, we were able to create a labelled dataset that can be used to identify spam game reviews in future research. Our method resulted in 5021 of the 33,450 unlabelled Steam reviews being labelled as spam reviews (about 15%). This falls within the expected range of 10-20% and maps to the Yelp figures of 14-20% of reviews are spam.

6 CONCLUSIONS

This paper reported on three initial projects that we have undertaken in order to investigate applying NLP techniques to the automatic analysis of video game reviews. We were successfully able to employ the Steam API to scrape very large sets of reviews across multiple games, along with associated metadata, in order to use in each project for testing with our various algorithms.

Each of our projects has provided initial data and insights from applying NLP techniques to game reviews for the purposes of feature summarisation, sentiment analysis, and identification of spam reviews. We identified both opportunities and challenges associated with applying each algorithm to our selected problems. There are many outstanding questions, including generalisability to different games, appropriate performance metrics, and non-English languages.

One common issue is a lack of labelled game review datasets to use for machine learning algorithms. We have created one such dataset for use in future research. We have also had some success in using labelled datasets from other domains (e.g., consumer product reviews), but they lack the specific context for games. Without labelled datasets specific to game reviews, we lack ground truth with appropriate context for understanding game related comments. Similarly, without labelled datasets, it is difficult to judge the accuracy of the algorithms. Valuable future research lies in creating labelled datasets of game reviews for use in research on automatic analysis of game reviews.

In conclusion, our results provide valuable initial steps in applying NLP techniques to analysing game reviews. Our goal is that, as a field, we will one day be able to automatically analyse large sets of game reviews to further understand player experience and expectation in video games, as well as to support developers in improving their games and players in making informed purchase decisions.

REFERENCES

- [1] Kushal Bafna and Durga Toshniwal. 2013. Feature based Summarization of Customers' Reviews of Online Products. *Procedia Computer Science* 22 (2013), 142 – 151. <https://doi.org/10.1016/j.procs.2013.09.090>
- [2] Farrah Benamara, Carmine Cesarano, Antonio Picariello, Diege R. Recupero, and Venkatramana S. Subrahmanian. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*. 1–7.
- [3] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.
- [4] Eduardo H. Calvillo-Gómez, Paul Cairns, and Anna L. Cox. 2015. *Assessing the Core Elements of the Gaming Experience*. Springer International Publishing, Cham, 37–62. https://doi.org/10.1007/978-3-319-15985-0_3
- [5] Claire Cardie. 2014. Sentiment Analysis and Opinion Mining. *Computational Linguistics* 40, 2 (2014), 511–513. https://doi.org/10.1162/COLI_r_00186
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:cs.CL/1810.04805*
- [7] L. Eberhard, P. Kasper, P. Koncar, and C. Gütl. 2018. Investigating Helpfulness of Video Game Reviews on the Steam Platform. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. 43–50.
- [8] Magdalini Eirinaki, Shamita Pital, and Japinder Singh. 2012. Feature-based opinion mining and ranking. *J. Comput. System Sci.* 78, 4 (2012), 1175 – 1184. <https://doi.org/10.1016/j.jcss.2011.10.007>
- [9] Geli Fei, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Exploiting Burstiness in Reviews for Review Spammer Detection. In *International AAAI Conference on Web and Social Media*. 175–184.
- [10] Jeffrey Fosset. 2017. *Yelp Summarization Miner (YUMM)*. Retrieved December 19, 2019 from <https://github.com/Fosset117/opinion-mining>
- [11] Steam Games. 2019. *Steamworks Documentation*. Retrieved December 19, 2019 from <https://partner.steamgames.com/doc/store/getreviews>
- [12] Ben Gifford. 2013. *Reviewing the critics: Examining popular video game reviews through a comparative content analysis*. Ph.D. Dissertation. Cleveland State University, Cleveland, OH.
- [13] Donato Hernández Fusilier, Rafael Guzmán Cabrera, Manuel Montes-y Gómez, and Paolo Rosso. 2013. Using PU-Learning to Detect Deceptive Opinion Spam. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Atlanta, Georgia, 38–45.

- [14] Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Seattle, WA, USA) (*KDD '04*). ACM, New York, NY, USA, 168–177. <https://doi.org/10.1145/1014052.1014073>
- [15] Clayton J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*.
- [16] Nitin Jindal and Bing Liu. 2008. Opinion Spam and Analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (Palo Alto, California, USA) (*WSDM '08*). ACM, New York, NY, USA, 219–230. <https://doi.org/10.1145/1341531.1341560>
- [17] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32* (Beijing, China) (*ICML '14*). JMLR.org, II–1188–II–1196.
- [18] Dayi Lin, Cor-Paul Bezemer, Ying Zou, and Ahmed E. Hassan. 2019. An empirical study of game reviews on the Steam platform. *Empirical Software Engineering* 24, 1 (2019), 170–207. <https://doi.org/10.1007/s10664-018-9627-4>
- [19] Li Luyang, Bing Qin, and Ting Liu. 2017. *2017 Survey on Fake Review Detection Research*. Research Center for Social Computing and Information Retrieval. Technical Report. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001.
- [20] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning* 39, 2 (2000), 103–134. <https://doi.org/10.1023/A:1007692713085>
- [21] Dim En Nyaung and Thin Lai Lai Thein. 2015. Feature-Based Summarizing and Ranking from Customer Reviews. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering* 9 (2015), 734–739.
- [22] Myle Ott, Claire Cardie, and Jeff Hancock. 2012. Estimating the Prevalence of Deception in Online Review Communities. In *Proceedings of the 21st International Conference on World Wide Web* (Lyon, France) (*WWW '12*). ACM, New York, NY, USA, 201–210. <https://doi.org/10.1145/2187836.2187864>
- [23] Myle Ott, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative Deceptive Opinion Spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, 497–501.
- [24] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 309–319.
- [25] Antoni Sobkowicz. 2017. *Steam Review Dataset*. Retrieved December 19, 2019 from <https://zenodo.org/record/1000885#.XbFasOgzaUk>
- [26] Steam Store. 2019. *Dota 2*. Retrieved December 19, 2019 from https://store.steampowered.com/app/570/Dota_2
- [27] Björn Strååt and Harko Verhagen. 2017. Using User Created Game Reviews for Sentiment Analysis: A Method for Researching User Attitudes. In *GHITALY@CHIItaly*.
- [28] Penelope Sweetser, Daniel Johnson, and Peta Wyeth. 2013. Revisiting the GameFlow Model with Detailed Heuristics. *The Journal of Creative Technologies* 3 (May 2013). <https://ojs.aut.ac.nz/journal-of-creative-technologies/index.php/JCT/article/view/16>
- [29] Penelope Sweetser, Daniel Johnson, Peta Wyeth, and Anne Ozdowska. 2012. GameFlow Heuristics for Designing and Evaluating Real-Time Strategy Games. In *Proceedings of The 8th Australasian Conference on Interactive Entertainment: Playing the System* (Auckland, New Zealand) (*IE '12*). ACM, New York, NY, USA, Article 1, 10 pages. <https://doi.org/10.1145/2336727.2336728>
- [30] Michal Trneny. 2017. *Machine learning for predicting success of video games*. Master's thesis. Masaryk University, Czech Republic.
- [31] Tom Wijman. 2019. *Newzoo's Games Trends to Watch in 2020*. Retrieved December 19, 2019 from <https://newzoo.com/insights/articles/newzoos-games-trends-to-watch-in-2020/>
- [32] Zhen Zuo. 2019. *Sentiment analysis of steam review datasets using naive bayes and decision tree classifier*. Technical Report. University of Illinois, Urbana-Champaign.