GA-Eval: A Neural Network Based Approach to Evaluate Video Games Acceptance

Augusto de Castro Vieira, Wladmir Cardoso Brandão
Department of Computer Science
Pontifical Catholic University of Minas Gerais (PUC Minas)
Belo Horizonte, Brazil
cv.augusto2@gmail.com, wladmir@pucminas.br

***Abstract*—Video game and interactive media are currently among the most profitable industries. In this competitive marketing, game developers are interested in designing products with aspects that increase user acceptance, such as a well written story, stable servers for multiplayer games, and fluid combat mechanics. Although user-expressed feelings about game aspects seem to correlate with user acceptance, sentiment analysis is under-exploited for video games user acceptance evaluation. In this article, we propose GA-Eval, an approach to evaluate the user acceptance of video games by using convolutional neural networks for aspect-based sentiment analysis of user text reviews. Experiments with open development games in the Early Access release model show that the sentiments expressed toward game aspects have strong correlations with game acceptance by users. In addition, experimental results show that convolutional neural networks are effective for the sentiment analysis of user reviews of video games.**

***Keywords*-video game, game acceptance, sentiment analysis, opinion mining, machine learning, neural network**

I. Introduction

Globally, video game and interactive media industries encompass one in three people as active consumers, generating an estimated revenue of 108.4 billion dollars in 2017 [1]. This huge user base and revenue have made them some of the largest and most competitive entertainment industries in the world. In this context, it is paramount for game and media producers to know which aspects of their products are relevant to their target consumers. Fortunately, these relevance feedback can be directly obtained from users and testers during the development and life cycles of the games. In particular, we can get feedback from user reviews provided on video games sales platforms such as *Steam*[1], which enable users to write critical reviews and assign positive or negative ratings for any game.

A recent work reported in literature show that the sentiment expressed by the users toward a given aspect of a game has a strong correlation to the user's acceptance of the game, but the process of collecting and evaluating reviewer's ratings is largely manual and mostly time-consuming [2]. In addition, automated sentiment analysis and opinion mining techniques are under-exploited for evaluating the acceptance of video games by consumers. In this article, we propose GA-Eval, an aspect-based sentiment analysis approach to evaluate the acceptance of video games by using convolutional neural networks (CNNs) to classify sentiments of user reviews of games in a complete Early Access cycle from *Steam* platform. As reported in literature, CNNs are an effective approach for automated sentiment analysis, particularly for sentiment classification [3], [4].

Different from a recent work reported in literature [2], we use a state-of-the-art sentiment classification approach [4] to perform aspect-based sentiment analysis. As a result, we automated and made the process of collecting and evaluating reviewer's ratings faster. Additionally, we run experiments using two video game datasets, and show that the sentiments expressed toward a popular aspect of a game have strong correlations to the user acceptance of the game. The main contributions of this article are: i) we propose an aspect-based sentiment analysis approach to evaluate the acceptance of video games by consumers; ii) we thoroughly evaluate our proposed approach attesting its effectiveness for automated sentiment analysis of user reviews on video games.

The remaining of this article is organized as follows: in the Section II, we review the related work reported in the literature on sentiment analysis of game reviews. In the Section III, we describe our approach to evaluate the acceptance of video games by consumers. In the Section IV, we present the experimental setup to evaluate our approach, including the dataset we collected from *Steam*, and the results of the experimental evaluation. Finally, in the Section V, we present our concluding remarks, as well as directions for future research.

II. Related Work

Recently, global vectors for word representation deep convolutional neural network (GloVe DCNN) was proposed for sentiment classification [4]. It uses CNNs and pre-trained word embeddings learned by the GloVe algorithm [5] in a corpus of *tweets*, as well as n-gram and sentiment polarity score features for expressing a *tweet*. The model provides classifications with accuracy ranging from 80% to 82%.

[1]*Steam* is a video game and software distribution platform owned and developed by Valve Corporation initially for releasing and selling their own games. However, it has grown over the years becoming the largest digital distribution platform for personal computer video games.

In the same line, an aspect-based sentiment analysis was proposed to evaluate the correlations between the sentiment classes positive, neutral, and negative and the user rating categories low, medium and high [2]. The authors collected user reviews from the *Metacritic* platform for six games belonging to the *Dragon Age* and *Mass Effect* franchises. The most frequent aspects were extracted using word frequency and only reviews with these aspects were used. A manual sentiment analysis was performed on the reviews and the correlations between the sentiment classes and the user rating categories were evaluated by a chi-square test of independence. The results show a strong correlation between user sentiments toward an aspect of a game and the game's ratings.

III. The GA-Eval Approach

GA-Eval is an aspect-based sentiment analysis approach to evaluate the acceptance of video games by consumers. Particularly, it can be used to investigate the existence of a correlation between the sentiment expressed toward the most popular aspects of a given game and the positive or negative user review ratings given a period of time. From these correlations, we can infer the level of user acceptance of a game and if the most frequently mentioned aspects impact user recommendations.

The first component of our approach is the crawler, which performs HTTP REST requests to the *Steamworks* API[2] to collect user reviews on games that had passed the complete Early Access cycle. The crawler component filters English reviews written or updated during the Early Access period and shortly after release. The criteria to select these periods is user activity measured by the average monthly quantity of concurrent players. The statistical data to support this selection is obtained from the *Steamcharts* platform[3].

The second component is transformer, that performs text preprocessing to prepare the collected reviews for aspects and features extraction. In particular, it performs eight preprocessing actions: i) words to lower case; ii) word separators to a single white space; iii) URLs removal; iv) numbers removal; v) acronyms expansion to their textual equivalents; vi) replacement of emoticons and emojis by their word equivalents; vii) removal of non-alphanumeric characters, since many acronyms, emoticons and emojis use such characters; viii) text tokenization.

The last three components of our approach, i.e., the extractor, the classifier and the evaluator, are presented in the sections III-A, III-B and III-C, respectively.

A. The Extractor Component

The extractor component extracts aspects and text features from user reviews. First, it performs text analysis to enumerate the most frequent words, since these words are associated with the most important aspects of a game in the minds of the reviewers [2]. The three most frequent aspects for each game are stored in the corpus of game aspects and features. Second, it extracts feature vectors to be used by the classifier component. The extracted feature vectors are essentially the same ones originally used by the GloVe DCNN model [4]: i) *n-gram:* for sentiment analysis, unigrams and bi-grams provide high performance in sentiment classification tasks on *tweets* [4]; ii) *sentiment polarity score*: the sentiment polarity of a word is the difference between the pointwise mutual information (PMI) of the negative and positive sentiments expressed by a word mapped into values from -5 (extreme negative) to 5 (extreme positive). We use the lexical dataset AFINN[4] to obtain the sentiment value for English words; iii) *word embeddings*: mapping of words into a vector space, where the distance between two vectors represents the semantic relationship between their correspondent words [4]. Strongly related words are represented by vectors that are close to one another in the vector space. We use GloVe [5] to learn the word embeddings of user reviews, training 100 dimensional word vectors using a well known implementation of GloVe[5].

B. The Classifier Component

The classifier component uses a CNN model for sentiment classification composed by one embedding layer and multiple convolution layers, following an architecture largely similar to [4], where the embedding layer corresponds to a concatenation of the word embeddings, sentiment polarity score and n-gram vectors. Following the convolutinal layers, we have a hidden layer consisting of a total of three convolution layers connected to three respective *k-max* pooling layers. Similarly to other research reported in literature [4], dropout regularization is performed to the fully connected layers to prevent over-fitting. The output layer is a *softmax* that generates the probabilities for three sentiment classes. The output layer uses a fully connected *softmax* to adjust the sentiment characteristics of the input layer, and provides the distribution probabilities for the sentiment classes.

C. The Evaluator Component

The evaluator component performs a chi-square independence test to confirm or refute the hypothesis of existence of correlation between the sentiments expressed by reviewers toward a game aspect and the user's propensity to recommend (or not recommend) the game, providing tests for the following hypotheses:

- **Null hypothesis**: there is no correlation between the sentiments expressed by reviewers toward a game aspect and their propensity to recommend the game;

[2] *Steamworks* is the official platform for partners interested in distributing their products through *Steam*. Available at http://partner.steamgames.com.

[3] *Steamcharts* provides statistical data of *Steam* games and users.

[4] http://www2.imm.dtu.dk/pubdb/p.php?6010.

[5] http://github.com/stanfordnlp/GloVe.

- **Alternative hypothesis**: there is a correlation between the sentiments expressed by reviewers toward a game aspect and their propensity to recommend the game.

The confirmation or refutation of the null hypothesis is determined by comparing the result of the chi-square test and the critical value corresponding to the degrees of freedom of it's contingency table. If the result of the chi-square test is above a critical value, the null hypothesis must be refuted, otherwise, it must be confirmed (accepted). In our approach we used a p-value of 0.01 and, since our contingency table has two degrees of freedom, the critical value of 9.21 must be exceeded for the null hypothesis to be refuted [6].

IV. EXPERIMENTS

To validate our evaluation approach, we run experiments to answer the following research questions: i) how effective is our approach to classify user reviews as positive, neutral, or negative? ii) is our aspect-based sentiment analysis approach a practical alternative to replace manual sentiment analysis; iii) are the most relevant game aspects mentioned by the reviewers correlated with their game recommendations? In the remainder of this section, we describe the experimental setup that supports these investigations.

A. Datasets

We use two datasets to run our experiments. The first one is reported in literature and comprises a sample of reviews from the *Dragon Age* franchise [2]. The second dataset[6] was collected by our evaluation approach as described in Section III. In particular, we collected user reviews from six games in the *Steam* platform that had passed the complete Early Access cycle: *ARK: Survival Evolved*, *Conan Exiles*, *Don't Starve Together*, *Rust*, *Subnautica*, and *The Forest*.

B. Setup

The training and validation of the GloVe DCNN model used by our evaluation approach were performed on a subset of reviews corresponding to 10% in both datasets, which were combined to manual sentiment analysis. Particularly, we train the model with the following hyper-parameters (initially set as [4] and tweaked during experimentation for more accurate results): i) *window sizes:* 1 of size 5 and 2 of size 7 for the convolution layers; ii) *# of filters:* 10 for each convolution layer; iii) *Dropout:* 0.5 for each convolution layer; iv) *# of hidden units:* 50 for each convolution layer; v) *Learning rate:* 0.001. The results of the sentiment classification were then distributed in contingency tables for the chi-square independence test.

[6]The dataset is available at https://doi.org/10.5281/zenodo.3403238

C. Experimental Results

1) Classification Effectiveness: We perform k-fold cross-validation with $k = 5$, using the accuracy of the sentiment classification as a measure of effectiveness. The result of these evaluation provides an average accuracy of 80% with a standard deviation of 1.64 and a confidence interval of 1.43 with a confidence level of 95%. Recalling our first research question, these observations attest the effectiveness of our approach to classify user reviews as positive, neutral or negative.

2) Automated Sentiment Analysis: We evaluate the feasibility of using our approach as a substitute for manual sentiment analysis by exploiting the *Dragon Age* dataset. In particular, we distribute the aspects, "Character", "Combat" and "Story" in contingency tables where each row represents a sentiment class for a review containing a given aspect, and each column represents the user ratings. Given that our CNN model classified a small amount of reviews as neutral, we performed the chi-square test both considering and not considering the neutral class. Table I shows the comparison of the results of the chi-square independence tests performed with the sentiment classes from the manual sentiment analysis with those from the aspect-based sentiment analysis, considering (and not) the neutral sentiment class.

Table I
CHI-SQUARE RESULTS: DRAGON AGE DATASET

Aspect	χ^2		
	Manual	**Aspect-based**	
		Neutral	**No Neutral**
Character	470.35	190.73	187.49
Combat	654.12	440.72	435.42
Story	923.37	414.48	409.34

From Table I, we can observe that the results of the chi-square test for both the manual and aspect-based sentiment analysis have led to the refutation of the null hypothesis, since both have exceeded the minimal critical value for the p-value of 0.01. Recalling our second research question, these observation attest that our aspect-based sentiment analysis approach is a practical alternative to replace manual sentiment analysis.

3) Evaluator Accuracy: Similarly to the test presented in Table I, we perform the chi-square independence test with and without the neutral sentiment class. In the later case, since we need to remove one row from the contingency table, the number of degrees of freedom became one, thus the critical value to be exceeded by the chi-square test result at p-value 0.01 is 6.635. Table II presents the results of the chi-square independence test for all games, aspects, and periods with enough review data necessary to perform the test. Recalling our third research question, these observation attest that our approach is effective to investigate the correlation between the most relevant game aspects mentioned by the reviewers and game recommendations.

Table II
CHI-SQUARE RESULTS FOR STEAM REVIEWS DATASET. RESULTS IN BOLD ARE BELOW THE CRITICAL VALUE

Game	Period	Aspect	χ^2 Neutral	χ^2 No Neutral
ARK : Survival Evolved	06/01/2015 - 09/30/2015	People	19.71	18.94
		Server	82.52	82.38
	12/01/2016 - 03/30/2017	Dinos	84.65	84.34
		People	61.51	60.88
		Server	84.81	84.75
	09/29/2017 - 10/01/2018	Dinos	141.52	138.65
		People	119.20	115.30
		Server	184.25	182.15
Conan Exiles	01/30/2017 - 05/30/2017	Bugs	51.72	51.56
		Server	93.14	92.87
	05/01/2018 - 08/30/2018	Bugs	78.79	78.58
		Combat	85.23	85.25
		Server	83.47	82.04
Rust	12/11/2013 - 04/30/2014	Friends	88.47	88.65
		People	109.67	109.81
		Server	76.85	76.68
	04/01/2015 - 09/30/2105	Friends	**1.04**	**0.29**
		People	**0.21**	**0.03**
		Server	**0.28**	**0.00**
	11/01/2015 - 03/01/2016	Friends	**2.03**	**0.48**
		People	**2.15**	**2.04**
		Server	**0.31**	**0.28**
	07/01/2016 - 10/30/2016	People	**1.29**	**0.91**
		Server	**0.02**	**0.01**
		System	**0.78**	**0.04**
	02/01/2018 - 06/01/2018	People	**1.29**	**0.52**
Subnautica	12/01/2017 - 04/01/2018	Exploration	152.83	150.33
		Story	54.44	50.70
		World	85.18	85.14
The Forest	04/30/2018 - 09/01/2018	Friends	**8.10**	7.35
		Story	16.39	15.10

V. CONCLUSION

In this article, we proposed GA-Eval, an aspect-based sentiment analysis approach to evaluate the acceptance of video games by users. Particularly, our evaluation approach collected user reviews of games from *Steam* platform and performed sentiment analysis in these reviews by using a convolutional neural networks to show that the sentiments expressed toward a popular aspect of a game have strong correlations to the user acceptance of the game. Experimental results attested the effectiveness of GA-Eval, with 80.00% of accuracy in sentiment classification. In addition, the chi-square test showed that our approach is a practical alternative to replace manual sentiment analysis and the results also showed that the most relevant aspects mentioned by the reviewers are strongly correlated with their recommendations.

In the future, we intend to address class imbalance and evaluate different strategies for sentiment analysis to improve the classification for the neutral sentiment class, starting with an in depth comparison of our model and other approaches reported in the literature.

ACKNOWLEDGMENT

This work was carried out with the support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Financing Code 001. The authors thank the support of the CNPq (Brazilian National Council for Scientific and Technological Development), FAPEMIG (Foundation for Research and Scientific and Technological Development of Minas Gerais), and PUC Minas.

REFERENCES

[1] SuperData, “2017 year in review: Digital games and interactive media,” SuperData Research Holdings Inc., Tech. Rep., 2018.

[2] B. Strååt and H. Verhagen, “Using user created game reviews for sentiment analysis: A method for researching user attitudes,” in *Proceedings of the 1st Workshop on Games-Human Interaction*, 2017.

[3] A. Severyn and A. Moschitti, “Twitter sentiment analysis with deep convolutional neural networks,” in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015, pp. 959–962.

[4] Z. Jianqiang, G. Xiaolin, and Z. Xuejun, “Deep convolution neural networks for Twitter sentiment analysis,” *IEEE Access*, vol. 6, pp. 23 253–23 260, 2018.

[5] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.

[6] W. F. Guthrie, *NIST/SEMATECH e-Handbook of Statistical Methods.* NIST, 2010.