A Step-by-Step Approach for User Acceptance Evaluation in Games Based on Sentiment Analysis and Machine Learning

Larissa F. S. Britto[1,2], Luciano D. S. Pacífico[3]

[1] Centro de Pesquisa e Desenvolvimento em Telecomunicações (CPQD), Brazil
[2] Centro de Informática (CIn), Universidade de Federal de Pernambuco (UFPE), Brazil
lfsb@cin.ufpe.br
[3] Departamento de Computação (DC), Universidade Federal Rural de Pernambuco (UFRPE), Brazil
luciano.pacifico@ufrpe.br

Abstract. User opinion analysis is an important tool to guide the decision-making process of independent game developers and game studios, once such activity leads to better product development towards the user satisfaction. Sentiment Analysis (SA) techniques have been widely used by companies to discover what customers are saying about their products, and the game industry could also benefit from such research field, once user feelings about a video game may have a relevant impact in retention and revenues. In this work, a thorough analysis on user acceptance in video games is performed by means of Natural Language Processing and SA approaches, where game reviews are exploited to understand what are the most relevant topics users are taking into consideration when they evaluate a game. A Sentiment Classification approach is performed, and the proposed methodology is discussed step-by-step, seeking out to motivate further researches in the field. Also, a new data set is proposed, composed by game reviews written in Brazilian Portuguese, given the relevance of Brazilian market in game industry.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Machine learning approaches**.

Keywords: Game Acceptance, Machine Learning, Natural Language Processing, Sentiment Analysis, Text Mining.

1. INTRODUCTION

With the significant improvement in communication technologies, such as mobile devices, in the past decades, video games have become one of the most profitable markets in entertainment industry. It is expected that the number of video game users will be 3.09 billion by the end of 2022, and the global market for video games will exceed US$203.1 billion at the and of this year[1]. To successfully operate in such a competitive market, where thousands of new games are released every year through multiple platforms, game developers need to understand the reasons and motivations that lead a game user to play their game, and, also, how to keep players engaged with their product. Game producers and developer teams need to understand what aspects of a game can increase user acceptance and engagement [Wang and Goh 2017; Vieira and Brandão 2019; Andreev et al. 2021].

With the ever growing popularity of the Internet and social networks, users can express their opinions freely, providing useful feedback and reviews concerning the games they liked (or disliked). Many game platforms, like Steam, allow users to provide reviews (written in natural language) and evaluations. Game reviews can be viewed as expert experience reports, as much as rich sources of user opinions and sentiments, giving players an idea on what to expect from the game, working as purchase guides to their readers [Livingston et al. 2011]. Sentiments expressed by the users about

[1] https://newzoo.com/insights/articles/games-market-revenues-will-pass-200-billion-for-the-first-time-in-2022-as-the-u-s-overtakes-china

Copyright©2022 Permission to copy without fee all or part of the material printed in KDMiLe is granted provided that the copies are not made or distributed for commercial advantage, and that notice is given that copying is by permission of the Sociedade Brasileira de Computação.

different aspects of the game have a strong correlation to the user's acceptance of that game [Strååt and Verhagen 2017]. In this context, Sentiment Analysis (SA), which is one of the most popular tasks in Natural Language Processing (NLP), attempts to explore text data automatically by the use of advanced Artificial Intelligence techniques, being useful to access user acceptance rates contained in game reviews, providing relevant information to guide future game design decisions and game improvements. Although useful, SA applications in game user acceptance evaluation is still a under-exploited theme, and just few works in that subject are available in the literature [Strååt and Verhagen 2017; Vieira and Brandão 2019; Andreev et al. 2021; Urriza and Clarino 2021].

As a manner to motivate more researches and the development of new applications to improve the understanding on Brazilian game users behavior (due to the fact that Brazil is a ever growing market for the game industry[2]), in this work an approach for video game user acceptance evaluation is proposed based on SA models. Once there is a shortage of resources and frameworks to promote SA applications in Brazilian Portuguese language, in an attempt to fill this gap, the main contributions of current proposal are: The development of a novel SA data set, based on game reviews in Brazilian Portuguese, for sentiment classification and game acceptance analysis; A step-by-step description of the proposed SA methodology for game users acceptance evaluation; A thorough Sentiment Classification evaluation using four well-established classifiers; A comparative experimental evaluation among different text feature extraction methods and text representation techniques; An analysis on the impact of stop words removal on the behavior of the selected classifiers; A complete qualitative and quantitative evaluation on the obtained results for the experimental evaluation.

The remainder of this work is organized as follows. The proposed methodology (including data set acquisition and preparation) is presented in Section 2. Our experimental setup and results are shown in Section 3, followed by some conclusions and leads to future works (Section 4).

2. METHODOLOGY

2.1 Proposed Data Set

The data set proposed in this work is composed by game reviews extracted from Steam[3] using its Web API. All collected reviews are in Brazilian Portuguese language, and they express what players think and how they feel about the games and their features. The data set will be available in a public repository[4] without any processing, giving the researchers the freedom to choose which pre-processing and feature extraction approaches are more suitable for their researches. The proposed data set preparation methodology is illustrated in Fig. 1.

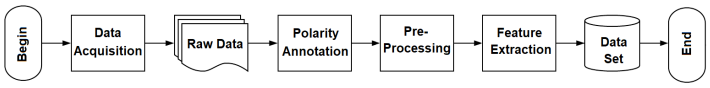
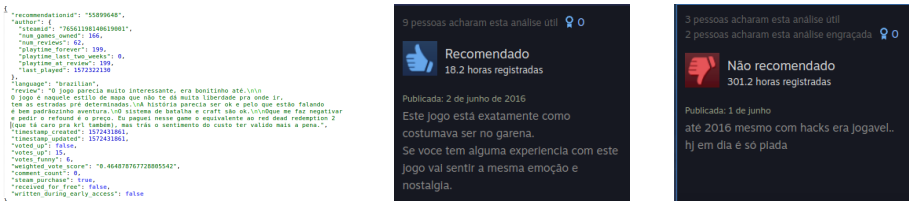


Fig. 1: Data set preparation steps.

The first step in data set creation is the acquisition of the data. In this work, Steam API[5] is employed for data acquisition. After the data acquisition, the sentiment polarity annotation is performed. All the reviews were classified in two classes, according to the game acceptance (*positive* or *negative*). The polarity is also obtained through the reviews made by each user. In Steam, the player can evaluate the game by recommending it or not (*vote_up* variable), so in the proposed data set, reviews in which the player recommends the game are considered with a **positive** polarity, and a **negative** polarity is

[2] https://www.statista.com/outlook/dmo/digital-media/video-games/brazil
[3] https://store.steampowered.com/
[4] https://github.com/larifeliciana/steam-reviews-portuguese
[5] https://steamcommunity.com/dev

attributed to reviews where the player does not recommend the game. An example of the information returned by our query in the Steam API, as much as samples for positive and negative evaluations are presented in Fig. 2.



(a) (b) (c)

Fig. 2: (a) Sample review returned by Steam API, (b) Review recommending a game, and (c) Review not recommending a game.

One of the most fundamental steps in text classification is the text document pre-processing, which includes data cleaning, which will remove irrelevant information and any other noise that may worsen the performance of the classifiers. Documents are also standardized. In this work, the following steps are performed for data pre-processing: **Lowercase Conversion**, where all document uppercase letters are converted into lowercase letters (such as “jogador"/“player"and “Jogador"/“Player") [HaCohen-Kerner et al. 2020]; **Stop Word Removal**, where words considered non-informative that occur with high frequencies in a document, like conjunctions, determiners and prepositions, are removed [Kaur and Buttar 2018]; **Special Character Removal**, where any special characters in the documents are removed, such as punctuation, symbols and digits, once they have no meaning and do not indicate any sentiment polarity, being totally disposable to sentiment classification purposes. Fig. 3 presents an example of the application of the pre-processing step in a document from the proposed data set.

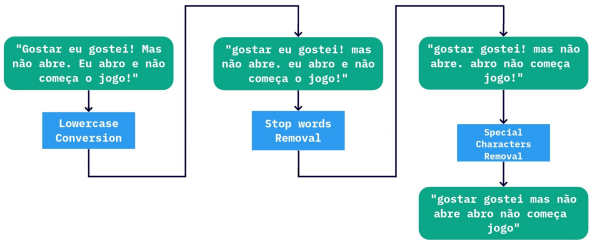


Fig. 3: The application of the pre-processing step in a document.

Feature extraction intends to transform raw documents from the data set into useful data supported by the classifiers. The following methods from the literature are adopted in this work:

(1) **Bag-of-Words** (BoW): A textual document is represented by its set of words. The text is converted into a matrix, where every column represents a word, each row represents a document, and each position contains the number of occurrences of a word in a document. The structure of the original document or the order of the words in that document are not taken into consideration by this representation [Kowsari et al. 2019]. The data set is also converted into a *Document-Term Matrix* (DTM).

(2) **Binary Bag-of-Words** (BBoW): Similar to standard BoW, but this time the DTM is a *binary matrix*, where each position indicates whether a given word belongs in or does not belong in the corresponding document.

Table I: Proposed data set stats.

Metric	Data Set	Positive Reviews	Negative Reviews
Total Number of Reviews	20000	10000	10000
Vocabulary Size	49587	37938	28765
Average Number of Words per Review	80.855	99.094	62.616
Average Number of Sentences per Review	4.1801	4.9214	3.4389
Average Number of Words per Sentence	18.722	18.715	18.729

(3) **Term Frequency–Inverse Document Frequency** (TF-IDF): TF-IDF measures how important a term is to a document in the data set, in order to lessen the effect of implicitly common words that occur with high frequencies, but have little relevance in the documents. The **Inverse Document Frequency** (IDF) has also been adopted as a feature extractor itself [Kowsari et al. 2019].

(4) **Delta TF-IDF** [Martineau and Finin 2009]: Instead of measuring how rare and important terms are, Delta TF-IDF measures how biased in *positive* or *negative* classes the terms are in the data set. Delta TF-IDF is calculated by the difference of the weight given by TF-IDF to the word in the positives and negatives documents.

3. EXPERIMENTAL EVALUATION

In this section, the experimental results obtained for the proposed data set are presented. Both quantitative and qualitative analysis are performed on the obtained results. The final statistics for data set are presented in Table I.

A five-folds cross-validation framework is employed in our evaluation, where the proposed data set has been randomly split into five balanced parts to form the training and testing sets. To generate a large variety of tests, the five-folds cross-validation process has been executed ten times, and, for each execution, five random distributions of the data have been obtained, in such a way that we had fifty different tests evaluations. The adopted resampling process has been performed to avoid results obtained by chance. Four well-known classification metrics are adopted: Accuracy, Precision, Recall and F-Measure. To execute the evaluation on the proposed data set, we compare the performances of four different and well-established classifiers from SA literature in sentiment classification task: Logistic Regression (LRC), Naive Bayes (NBC), Random Forest (RFC) and Support Vector Machines (SVM) [Kowsari et al. 2019]. We also compare the influence of five feature extraction methods for textual data (when applied to Brazilian Portuguese documents) on the behavior of the selected classifiers, as well as the influence of stop words removal, and the use of n-grams (*unigrams* and *bigrams*) to represent the relationships between words in the documents. Each one of the selected classifiers has its own set of hyper-parameters, and to find the best hyper-parameter configuration for these classifiers, a grid search is performed using the Machine Learning package Scikit-Learn [Pedregosa et al. 2011]. The search space for the hyper-parameters tuning task (based on [Yang and Shami 2020]), as well as the best final set of hyper-parameters, are presented in Table II. The hyper-parameters tuning task takes into consideration a different set of documents, composed by 4,000 reviews (2,000 “positive"reviews and 2,000 “negative"reviews). We employed the ten times 5-folds cross-validation approach for this task, and the obtained best results are related to an empirical analysis concerning the average testing accuracy.

After hyper-parameters tuning, the quantitative evaluation is performed considering an empirical analysis on the average testing values for all four classification metrics, and an overall rank system employed through the application of Friedman-Nemenyi non-parametric hypothesis test on the testing accuracies [Demšar 2006]. Since the testing accuracy is a *maximization metric*, the best algorithms will find higher average ranks for the Friedman-Nemenyi test. The experimental results are shown in Table III.

Considering an empirical analysis, we can observe that the best classifiers are able to obtain average accuracies over 0.89, what can be considered a good sentiment classification rate. Table III shows that

Table II: Configuration space for the hyper-parameters of tested Machine Learning models. In bold, the best hyper-parameters configuration.

Classifier	Hyper-Parameters	Search Space
LRC	C	0.001, 0.01, 0.1, **1**, 10
	penalty	l1, **l2**, elasticnet, none
	solver	newton-cg, lbfgs, **liblinear**, sag, saga
NBC	alpha	0.001, 0.01, 0.1, **1**, 10
RFC	criterion	gini, **entropy**
	max_features	**1**, 64
	min_samples_leaf	**1**, 11
	min_samples_split	2, **11**
	max_depth	5, 50, 100, **none**
	n_estimators	10, 100, 200, 300, **500**
SVM	C	0.001, 0.01, **0.1**, 1, 10
	penalty	l1, **l2**

Table III: Experimental results. SW: stop words are present in the documents; SWR: stop words are removed from the documents. U: only unigrams are used; UB: both unigrams and bigrams are used.

Bag-of-Words (BoW)

Algorithm	*n*-grams	Accuracy		Precision		Recall		F-measure		Execution Time	
		SW	SWR	SW	SWR	SW	SWR	SW	SWR	SW	SWR
LRC	U	0.8822	0.8794	0.8815	0.8791	0.8833	0.8799	0.8823	0.8799	2.0940	1.6043
	UB	**0.8954**	0.8931	**0.8981**	0.8942	0.8921	0.8920	0.8951	0.8920	7.0985	5.4894
NBC	U	0.8466	0.8502	0.8720	0.8724	0.8126	0.8204	0.8412	0.8204	1.1091	**1.0223**
	UB	0.8658	0.8696	0.8435	0.8481	0.8983	**0.9006**	0.8700	**0.9006**	3.5847	3.2290
RFC	U	0.8638	0.8701	0.8569	0.8603	0.8737	0.8837	0.8652	0.8837	74.3529	80.4900
	UB	0.8783	0.8846	0.8700	0.8773	0.8896	0.8945	0.8796	0.8945	322.1164	370.7713
SVM	U	0.8799	0.8779	0.8799	0.8775	0.8800	0.8786	0.8799	0.8786	2.3931	2.0363
	UB	0.8916	0.8889	0.8935	0.8899	0.8893	0.8876	0.8913	0.8876	19.4711	16.8304

Binary Bag-of-Words (BBoW)

Algorithm	*n*-grams	Accuracy		Precision		Recall		F-measure		Execution Time	
		SW	SWR	SW	SWR	SW	SWR	SW	SWR	SW	SWR
LRC	U	0.8790	0.8788	0.8789	0.8756	0.8793	0.8833	0.8790	0.8833	1.5612	1.4190
	UB	**0.8930**	0.8902	**0.8948**	0.8891	0.8908	0.8918	0.8928	0.8918	5.8213	4.7434
NBC	U	0.8390	0.8427	0.8774	0.8696	0.7882	0.8064	0.8303	0.8064	1.0801	**1.0239**
	UB	0.8701	0.8717	0.8596	0.8555	0.8849	**0.8947**	0.8720	**0.8947**	3.6044	3.2180
RFC	U	0.8606	0.8676	0.8555	0.8590	0.8679	0.8799	0.8616	0.8799	72.2700	80.4281
	UB	0.8770	0.8844	0.8719	0.8783	0.8840	0.8925	0.8779	0.8925	328.9788	373.4355
SVM	U	0.8772	0.8761	0.8770	0.8741	0.8776	0.8790	0.8772	0.8790	1.3709	1.2595
	UB	0.8894	0.8868	0.8902	0.8850	0.8884	0.8891	0.8892	0.8891	12.0697	9.1356

Delta TF-IDF

Algorithm	*n*-grams	Accuracy		Precision		Recall		F-measure		Execution Time	
		SW	SWR	SW	SWR	SW	SWR	SW	SWR	SW	SWR
LRC	U	0.8852	0.8864	0.8851	0.8853	0.8855	0.8879	0.8853	0.8879	1.9971	1.7061
	UB	0.8885	0.8882	0.8865	0.8809	0.8913	0.8980	0.8889	0.8980	6.1673	5.0962
NBC	U	0.8596	0.8609	0.8324	0.8382	0.9009	0.8945	0.8652	0.8945	1.8994	**1.6134**
	UB	0.8590	0.8733	0.8117	0.8379	**0.9352**	0.9258	0.8690	**0.9258**	5.8268	4.8320
RFC	U	0.8640	0.8706	0.8704	0.8728	0.8556	0.8679	0.8629	0.8679	76.3051	80.8928
	UB	0.8767	0.8819	0.8827	0.8918	0.8689	0.8695	0.8757	0.8695	302.0472	339.1312
SVM	U	0.8878	0.8882	0.8883	0.8879	0.8873	0.8887	0.8877	0.8887	1.9898	1.6996
	UB	**0.8920**	0.8914	**0.8908**	0.8847	0.8936	0.9002	0.8921	0.9002	6.1968	5.1288

IDF

Algorithm	*n*-grams	Accuracy		Precision		Recall		F-measure		Execution Time	
		SW	SWR	SW	SWR	SW	SWR	SW	SWR	SW	SWR
LRC	U	0.8842	0.8839	0.8825	0.8813	0.8865	0.8874	0.8845	0.8874	1.2088	1.1346
	UB	0.8901	0.8885	0.8906	0.8821	0.8894	0.8970	0.8900	0.8970	4.1724	3.7015
NBC	U	0.8584	0.8580	0.8377	0.8360	0.8892	0.8908	0.8626	0.8908	1.0702	**1.0483**
	UB	0.8646	0.8736	0.8236	0.8407	**0.9282**	0.9219	0.8727	**0.9219**	3.8592	3.4574
RFC	U	0.8587	0.8629	0.8571	0.8556	0.8610	0.8733	0.8590	0.8733	76.2363	81.6152
	UB	0.8746	0.8800	0.8719	0.8796	0.8782	0.8806	0.8750	0.8806	302.0712	339.4197
SVM	U	0.8874	0.8863	0.8863	0.8839	0.8888	0.8896	0.8875	0.8896	1.1990	1.1268
	UB	**0.8921**	0.8904	**0.8932**	0.8841	0.8907	0.8986	0.8919	0.8986	4.1513	3.7568

TF-IDF

Algorithm	*n*-grams	Accuracy		Precision		Recall		F-measure		Execution Time	
		SW	SWR	SW	SWR	SW	SWR	SW	SWR	SW	SWR
LRC	U	0.8852	0.8849	0.8859	0.8847	0.8843	0.8852	0.8851	0.8852	1.2002	1.1420
	UB	0.8874	0.8864	0.8886	0.8828	0.8860	0.8912	0.8872	0.8912	4.2016	3.7013
NBC	U	0.8584	0.8595	0.8332	0.8373	0.8963	0.8924	0.8636	0.8924	1.0693	**1.0516**
	UB	0.8589	0.8709	0.8125	0.8355	**0.9332**	0.9237	0.8686	**0.9237**	3.8612	3.4707
RFC	U	0.8632	0.8685	0.8612	0.8628	0.8659	0.8765	0.8635	0.8765	75.4396	81.6579
	UB	0.8767	0.8819	0.8740	0.8825	0.8804	0.8811	0.8771	0.8811	301.6509	338.0997
SVM	U	0.8876	0.8884	0.8886	0.8885	0.8864	0.8883	0.8875	0.8883	1.1938	1.1314
	UB	**0.8906**	0.8892	**0.8931**	0.8856	0.8876	0.8939	0.8903	0.8939	4.1800	3.7184

all algorithms have not been considerably affected by the stop words removal approach, considering all five feature extraction methods, in terms of the average values for the selected classification metrics. But, although the average values are not significantly affected, a larger feature space have increased the average execution time for almost all evaluated classifiers, except for Random Forest. Once RFC is an ensemble of Decision Trees, the increasing in the execution time after stop words removal may be related to the fact that each estimator may have needed more complex decision rules (i.e., deeper Decision Trees) to compensate the reduced feature space, affecting the execution of the algorithm. By the other hand, when the evaluation takes into consideration the representation of the relationships between words in the documents by *n*-grams, all algorithms have been affected in some degree, and, in most cases, the algorithms presented a better performances when both unigrams and bigrams are adopted. But the impact on the average execution time is quite relevant in most cases (for instance, for RFC, considering TF-IDF feature extraction approach and stop words removal, the average execution

Table IV: Overall evaluation: average ranks for the Friedman-Nemenyi Test for accuracy metric, with $CD = 10.3161$. **Bold**: the best average rank corresponding to a feature extraction method.

Algorithm	BoW	BBoW	Delta TF-IDF	IDF	TF-IDF
LRC$_{U,SW}$	97.9700	90.5300	102.6700	101.7400	104.6900
LRC$_{UB,SW}$	**139.8800**	**139.0400**	115.3800	**123.6600**	114.3000
LRC$_{U,SWR}$	88.6200	89.5600	108.2800	101.4300	104.1000
LRC$_{UB,SWR}$	**139.8800**	**139.0400**	115.3800	**123.6600**	114.3000
NBC$_{U,SW}$	10.2000	9.4500	21.3500	22.0800	21.7400
NBC$_{UB,SW}$	40.5600	56.8200	21.6200	37.7400	23.9600
NBC$_{U,SWR}$	14.0500	12.5400	24.8800	21.7100	23.9800
NBC$_{UB,SWR}$	54.2300	63.6100	57.6600	64.4600	53.9700
RFC$_{U,SW}$	36.4600	34.2000	30.0700	24.1100	33.5900
RFC$_{UB,SW}$	82.5600	81.8000	68.1600	67.0000	71.8100
RFC$_{U,SWR}$	55.6500	48.2100	49.3900	33.0500	46.8800
RFC$_{UB,SWR}$	107.3500	111.8200	88.8900	86.3400	91.7500
SVM$_{U,SW}$	89.8200	83.5800	113.0800	114.8800	115.1100
SVM$_{UB,SW}$	128.5300	127.8100	**128.8000**	**131.4200**	**127.3800**
SVM$_{U,SWR}$	81.5200	80.3300	115.6100	110.3500	**118.4500**
SVM$_{UB,SWR}$	120.7200	119.6600	**126.7800**	**124.3700**	**121.9900**

time when only unigrams are considered is about 82 seconds, but when both unigrams and bigrams are adopted, the average execution time reaches 338 seconds), what is completely expected, once the adoption of bigrams increases significantly the problem feature space, so the designers of a sentiment classification system should take this trade-off into consideration when projecting their systems. The classifier that showed the higher degree of instability was the NBC, with average accuracies ranging from 0.839 (with Bag-of-Words, no stop words removal, and considering unigrams only) to 0.8736 (with IDF, stop words removal, and considering both unigrams and bigrams). RFC presented almost the same degree of stability when considering different feature extraction process, and both LRC and NBC have presented more instability when BoW and BBoW are used (Bag-of-Words-based methods), instead of Delta TF-IDF, IDF and TF-IDF (TF-IDF-based methods). That is also expected once TF-IDF, IDF and TF-IDF feature weighting approaches are able find the most relevant words for each document. SVM presented the higher stability considering different feature extraction methods from the evaluated classifiers, and both SVM and LRC have been able to find the best average values in our empirical analysis.

The overall evaluation represented by the ranking system obtained by the application of Friedman-Nemenyi hypothesis test on accuracy metric for all classifiers and feature extraction methods is presented in Table IV. The ranking system corroborates that stop words removal has a lesser impact on the behavior of the selected classifiers than the adopted document representation by n-grams. Logistic Regression has been able to obtain the best overall accuracy value (with significantly statistical differences) in relation to all other comparison classifiers, for both BoW and BBoW, but for Delta TF-IDF and TF-IDF, the best overall accuracy has been achieved by SVM. Although SVM has achieved slight better ranks for IDF feature extraction method than LRC, the Friedman-Nemenyi hypothesis test has not found significantly statistical differences between both approaches, so they are considered equivalent in this case.

(a) (b)

Fig. 4: Most frequent terms in positive reviews (a) Unigrams (b) Bigrams.

(a) (b)

Fig. 5: Most frequent terms in negative reviews (a) Unigrams (b) Bigrams.

The evaluation on what users are saying about a game in their reviews is another great way to understand game acceptance by customers. By means of our proposed qualitative analysis, we are

Table V: Top unigrams and bigrams selected by each feature extraction method.

BoW		BBoW		Delta TF-IDF		IDF		TF-IDF	
Unigram	Bigram	Unigram	Bigram	Unigram	Bigram	Unigram	Bigram	Unigram	Bigram
jogo	jogo não	jogo	jogo não	jogo	jogo bom	jogo	jogo bom	jogo	jogo bom
não	jogo bom	não	jogo bom	não	jogo não	não	jogo não	não	jogo não
mas	mas não	mas	vale pena	mas	vale pena	bom	vale pena	bom	vale pena
pra	vale pena	bom	mas não	bom	não recomendo	mas	não recomendo	mas	bom jogo
bem	não recomendo	jogar	não recomendo	pra	mas não	jogar	mas não	pra	não recomendo
jogar	trilha sonora	pra	trilha sonora	jogar	bom jogo	recomendo	bom jogo	jogar	melhor jogo
bom	desse jogo	bem	pra jogar	bem	melhor jogo	pra	melhor jogo	recomendo	mas não
game	pra jogar	recomendo	desse jogo	recomendo	pra jogar	melhor	pra jogar	bem	jogar amigos
vai	não comprem	melhor	bom jogo	melhor	jogar amigos	bem	jogar amigos	melhor	não gostei
história	jogo bem	ainda	jogo bem	game	não vale	divertido	não vale	game	pra jogar

able to know what players consider to be the good or bad characteristics of a game, and, also, what game features are well-accepted or not. Figures 4 and 5 present, respectively, the most frequent terms (in Brazilian Portuguese) in positive and negative reviews. In both positive and negative reviews, there is a predominance of neutral terms, meaningless in Sentiment Analysis context, such as “jogo"(“game"), “jogar"(“play"), “não"(“no"), “desse jogo"(“this game"). Terms that express users opinion or feelings are also very frequent, like “pior"(“worst"), “não comprem"(“don’t buy it"), “não recomendo"(“not recommended") in negative reviews, and “ótimo"(“great"), “perfeito"(“perfect"), “jogo bom"(“good game") in positive reviews. Looking closely, we are able to mining more meaningful information, as the points about the game players consider positve or negative, which is very relevant information for game companies and developers to know what problems must be solved or what users think that should be improved. Among the most positively discussed points are graphics (“gráficos bons"/ “good graphics"), story (“boa história"/ “good story", or “gostar história"/ “to like story"), soundtrack (“trilha sonora"/ “soundtrack", “funk brasileiro"/ “brazilian funk") and gameplay (“boa jogabilidade"/ “good gameplay"). The main negative point in games are errors and bad game performance (“muitos bugs"/ “many bugs", “alguns bugs"/“some bugs", “jogo crasha"/ “the game crashes", “mal otimizado"/ “poorly optimized"). To evaluate the feature balancing capabilities of the selected feature extraction methods, in selecting and balancing the text features from the proposed data set, the top-10 unigrams and bigrams extracted with each method are presented in Table V. Although both methods have selected many common terms, Bag-of-Words-based approaches (BoW and BBoW) gave high weights to irrelevant features, as “jogo não"(“game don’t"), “mas não"(“but not"), “pra"(“to"). The methods that derived from TF-IDF (Delta TF-IDF, IDF and TF-IDF) have been able to make a better ranking, giving high weights to most significant terms, such as “vale pena"(“it’s worth"), “não recomendo"(“not recommended"), “bom"(“good"), and also, such approaches have been able to select some important words that have not been selected by Bag-of-Words-based methods, such as “recomendo"(“recommended"), “melhor"(“best"), “melhor jogo"(“best game") e “não gostei"(“I didn’t like"), demonstrating a better capacity for feature selection and weighting.

4. CONCLUSION

In this work, video game acceptance was evaluate through a sentiment polarity classification approach, based on the fact that sentiments expressed in reviews about game aspects are strongly correlated to the user acceptance of the game. By classifying the game reviews, we can infer the user acceptance about a game. A sentiment data set in the domain of games, composed by reviews in Brazilian Portuguese, was proposed. The proposed data set was analyzed in an attempt to understand what topics and elements from the games are more relevant to the players, positively and negatively, and which ones are not. We evaluate the performance of four well-established classifiers from Sentiment Analysis and Machine Learning literature: Logistic Regression, Naive Bayes, Random Forest and Support Vector Machines. We also evaluated the influence of five text feature extraction methods, stop words removal, and two different document representation with n-grams approaches. The experimental results pointed out that Support Vector Machines and Logistic Regression are able to achieve the best performances in relation to the four selected classification metrics (Accuracy, Precision, Recall and F-Measure). Logistic Regression has also achieved the best average execution time. The experimental evaluation also showed that, for the proposed data set, the selected approach for relationship representation between words using n-grams and the feature extraction methods represented higher

impacts on the performances of the selected classifiers than stop words removal approach, although stop words removal may represent a reduction on the average execution time (as much as a reduction on storage memory) demanded by most of the selected classifiers. The best overall performances for the proposed data set have been achieved by SVM and LRC classifiers. By analyzing the most frequent words, we could have a better understanding on how players feel about a game, as much as what they are discussing, what are their favorite game elements, and how good their experiences with that game have been. In positive reviews, we can observe some well-accepted elements, such as graphics, gameplay, soundtrack and storytelling. On the other hand, problems concerning the game performance and bugs are the most frequently reported in negative comments. The use of SA techniques allows the identification of problems in the game, areas of improvement and the understanding on what to change to increase video game acceptance. In the future, we intend to extend this work by analyzing how the game acceptance varies with time, as an attempt to understand how game changes and updates can influence the game acceptance by its users. Also, we intend to adopt quantitative analysis, provided by Game Analytics and User Modeling techniques, to complement the qualitative analysis of SA, so we could have a better understanding on player's interaction and engagement with the games.

Acknowledgments

The authors would like to thank CPQD for the financial support.

REFERENCES

Andreev, G., Saxena, D., and Verma, J. K. Impact of review sentiment and magnitude on customers' recommendations for video games. In *2021 International Conference on Computational Performance Evaluation (ComPE)*. IEEE, pp. 992–995, 2021.

Demšar, J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research* vol. 7, pp. 1–30, 2006.

HaCohen-Kerner, Y., Miller, D., and Yigal, Y. The influence of preprocessing on text classification using a bag-of-words representation. *PLOS ONE* 15 (5): 1–22, 05, 2020.

Kaur, J. and Buttar, P. K. A systematic review on stopword removal algorithms. *Int. J. Futur. Revolut. Comput. Sci. Commun. Eng* 4 (4), 2018.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., and Brown, D. E. Text classification algorithms: A survey. *CoRR* vol. abs/1904.08067, 2019.

Livingston, I., Nacke, L., and Mandryk, R. The impact of negative game reviews and user comments on player experience. In *Proceedings of the 2011 ACM SIGGRAPH Symposium on Video Games*. ACM, pp. 25–29, 2011.

Martineau, J. and Finin, T. Delta tfidf: An improved feature space for sentiment analysis. In *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 3. pp. 1–4, 2009.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* vol. 12, pp. 2825–2830, 2011.

Stråt, B. and Verhagen, H. Using user created game reviews for sentiment analysis: A method for researching user attitudes. In *GHITALY17: 1st Workshop on Games-Human Interaction*. ACM, pp. 1–6, 2017.

Urriza, I. M. and Clarino, M. A. A. Aspect-based sentiment analysis of user created game reviews. In *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. pp. 76–81, 2021.

Vieira, A. C. and Brandão, W. C. GA-Eval: a neural network based approach to evaluate video games acceptance. In *Proceedings of the 18th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames 2019)*. SBC, pp. 595–598, 2019.

Wang, X. and Goh, D. Video game acceptance: A meta-analysis of the extended technology acceptance model. *Cyberpsychology, Behavior, and Social Networking* vol. 20, pp. 662–671, 11, 2017.

Yang, L. and Shami, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing* vol. 415, pp. 295–316, 2020.