SteamBR: a dataset for game reviews and evaluation of a state-of-the-art method for helpfulness prediction

Germano A. Z. Jorge, Thiago A. S. Pardo

Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of Sao Paulo
Av. Trabalhador São Carlense, 400 – 13.566-590 – São Carlos – SP

germano.jorge@usp.br, taspardo@icmc.usp.br

Abstract. *The digital revolution has led to exponential growth in user-generated content, including ratings and reviews, across numerous online platforms. One such platform is Steam, a multifaceted digital distribution network primarily for video games, that also functions as an active social network. Like many e-commerce, travel, and restaurant platforms, Steam users rely heavily on reviews to inform their purchasing decisions. However, the vast amount of data and varying quality of reviews may hinder the utility of such reviews. Furthermore, there is a significant challenge in assessing the helpfulness of recent or less-voted reviews. This study proposes a method for automating review helpfulness evaluation, focusing particularly on Brazilian Portuguese game reviews. The research involved the collection of a large dataset, including 2,789,893 reviews from over 12,000 games, creating a novel dataset for game reviews. Using feature extraction techniques, we were able to capture the metadata, semantic elements, and distributional characteristics present in the reviews. Subsequently, Machine Learning algorithms were employed to perform classification and regression tasks, with the objective of discerning helpful from unhelpful reviews. The achieved results demonstrated that the method was highly effective in predicting review helpfulness.*

1. Introduction

The popularization and increasing use of the web have made vast the amount of ratings and reviews available on online sites, such as those for e-commerce, movies, travel, or games. In this way, a consumer can rely on various opinions from others to become more informed about what they want and feel safer before making their purchase [Bertaglia 2017, Krishanamoorthy 2015, Sousa et al. 2019, Zhang et al. 2006]. However, this large number of reviews also includes those of dubious quality, vague opinions, and poorly written text [Kim et al. 2006]. Thus, the extensive amount of data and unwanted reviews can be a significant difficulty for users.

On most of these sites, a user can rate a review as helpful or not, so the most helpful ones are at the top. The helpfulness rating can be of great value to the consumer. Positive reviews can help in the verification of attributes and features that the user considers essential in the product and/or service and influences the final purchase decision. Interestingly, the helpfulness of reviews also helps in combating spam and fake content (such relevant topics nowadays), as it downgrades negatively rated reviews [Liu 2012].

In this context, aiming to advance research on this front, the present work proposes the

creation of SteamBR, a new large dataset in Portuguese in a new domain (games). It also explores a state-of-the-art method to automate the evaluation of the helpfulness of reviews, relying on a supervised machine learning model with linear regression and a classification model. The remainder of this paper is organized as follows. Section 2 introduces some essential concepts and the main related work. Section 3 contains a detailed description of the dataset built and used. In Section 4, experiments and results are described. Finally, Section 5 makes concluding remarks.

2. Related work

In one of the main works for Portuguese, Sousa et al. [2019] performed experiments aiming to detect the helpfulness and bias of user reviews in Apps and Movies using the UTLcorpus dataset. The highest F1 value was obtained using the Multi-Layer Perceptron (MLP) model, reaching 0.66 for utility prediction and 0.74 for polarity classification. For the same dataset, using a Convolutional Neural Network model, Sousa and Pardo [2022] achieved an F1 of 0.9 for apps and 0.74 for movies. Another type of learning model was proposed by Barbosa et al. [2016], who used an automated method using an artificial neural network to analyze the helpfulness of user reviews on Steam and verify which features made a review helpful. Performance was evaluated using k-fold cross-validation, and the RMSE found was 0.1929, considered acceptable by the authors. In this current article, we drew upon the study conducted by [Baowaly et al. 2019] as a primary reference. The authors collected English reviews of games on Steam and proposed a classification model that consisted of two prediction tasks: one classification and one regression. They employed the GBM (Gradient Boosting Machine) algorithm to train the model and selected both textual and metadata features. Results comparing our proposed model to these ones are displayed in Section 4.

3. Dataset Creation

The data used to build the working dataset was collected through a Web Crawler [Abukausar et al. 2013] on the Steam gaming platform, which analyzed 12,872 games and generated a JSON file for each containing its ID and its reviews. Additionally, crucial metadata information was obtained, such as the number of votes and the review author's recommendation regarding the game.

Individual files were systematically assigned a unique Steam identifier, such as `976310`, for example. To ascertain the name and genre of each corresponding game, a human annotator was required to manually enter this ID into the internet's browser address bar. It is important to note that certain games were associated with multiple genres. In these cases, the annotator had extensive gaming experience to select the genre that most accurately represented the game's primary attributes. Finally, the files were subsequently arranged following this pattern: `Genre_NameOfGame`. This systematic categorization allowed for a clearer representation of each game's primary genre and title. The decision to categorize games by genre was based on the work of Baowaly et al. [2019], which highlights that this is an essential step in training utility prediction algorithms.

Table 1 presents the number of reviews by genre. In total, 2,789,893 reviews were collected, of which only those with votes >=3 were selected, totaling 233,824 reviews divided

into ten different genres. We choose only reviews with more than 3 votes because ratings with a higher number of votes tend to reflect a more accurate assessment [Baowaly et al. 2019]. Moreover, it's crucial to emphasize that reviews without any votes cannot be conclusively determined to have been viewed. The absence of votes may suggest limited visibility or engagement, underscoring the complexity of interpreting user interactions.

Table 1. Number of reviews per game genre

Genre	Raw Dataset number of reviews	Filtered Dataset (at least three votes) number of reviews
Action	734,894	65,164
Indie	504,648	39,442
RPG	469,548	38,658
Adventure	366,078	33,088
Strategy	189,073	17,842
Simulation	157,536	11,164
Horror	148,510	12,880
FPS	102,368	7,714
Racing	93,743	7,166
Sports	23,495	1,966
Total	**2,789,893**	**233,824**

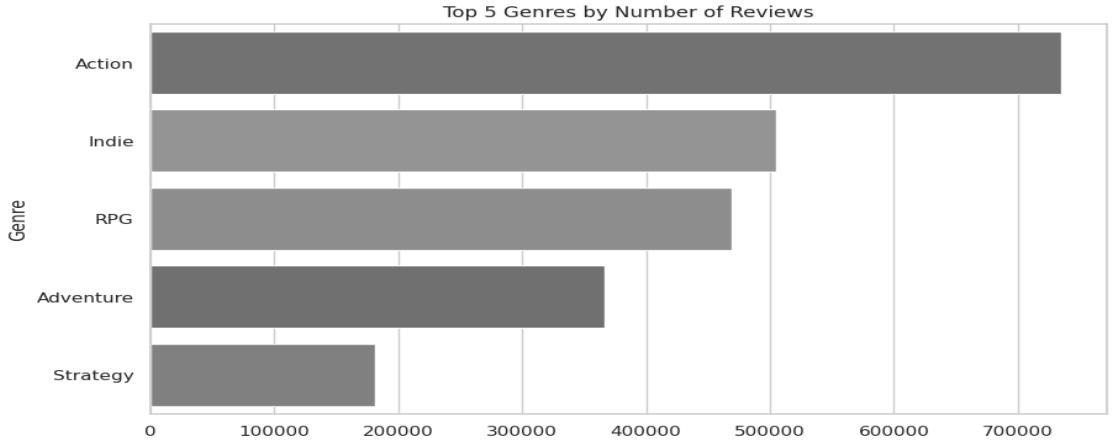

Figure 1. Genre-Wise Review Distribution

Figure 1 illustrates the discourse around gaming genres. Action games are the most reviewed at 700,000, reflecting high player engagement. Indie games, with 500,000 reviews, show significant discussion around their diverse experiences. Adventure games have 350,000 reviews, suggesting keen interest in narratives. Strategy games, despite their diversity, have only 1,900 reviews.

Next, Figure 2 offers a valuable snapshot of the game genre distribution on the platform. Indie games, with over 3,000 titles, stand as the most abundant genre, reflecting a vibrant independent game development community. Action and Adventure games, amounting to approximately 2,500 and 1,700 titles respectively, demonstrate a preference for engaging and explorative gameplay. Other genres like RPG, Strategy, and Simulation

have a solid presence, while Horror, Racing, FPS, and Sports are less common.

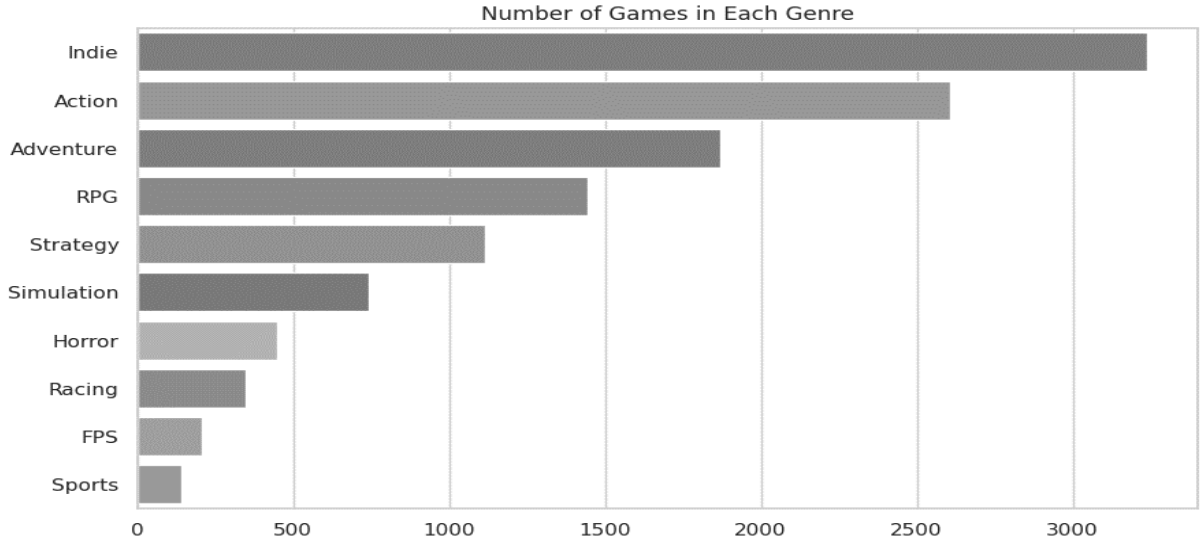

Figure 2. Distribution of Games Across Genres

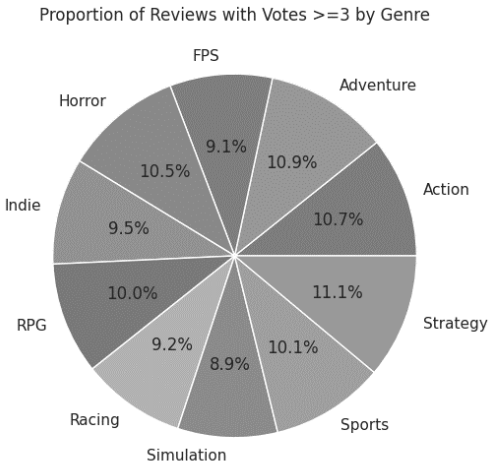

Figure 3. The proportion of Reviews by Genre with votes >=3

The results in Figure 3 showed a consistent pattern, with about 10% of reviews per genre receiving three or more votes. This indicates that genre is not a significant factor in a review engagement, suggesting that other aspects like review quality, posting time, or game popularity might be more influential in determining community interaction with a review.

4. Experiments and Results

Following the steps of Baowaly et al. [2019], only reviews with 3 votes or more were used. The reviews were pre-processed and features of different types were extracted, such as metadata [Baowaly et al. 2019, Liu et al. 2007, Lu et al. 2010, Kim et al. 2006, Mudambi and Schuff 2010], sentiment [Kim et al. 2006, Balage Filho et al. 2013, Pennebaker et al. 2001], topics

[Baowaly et al. 2019, Blei et al. 2017] and embeddings [Mikolov et al. 2013, Le & Mikolov 2014]. We then applied the GBM algorithm [Friedman 2001] for pattern recognition.

Table 2. Main results from 4 different studies in helpfulness prediction.

	F1-Scores			RMSE		
Category	Baowaly et al (2019)	Jorge & Pardo (2023)	Sousa & Pardo (2022)	Baowaly et al (2019)	Jorge & Pardo (2023)	Barbosa et al (2016)
Combined	0.94	0.90	-	0.170	0.095	0.1929
Apps	-	-	0.9	-	-	-
Movies	-	-	0.74	-	-	-

Our method has shown strong performance. The F1-score was 0.90, marginally below Baowaly et al. [2019]'s 0.94. Table 2 shows our study's Root Mean Square Error (RMSE) values, indicative of model fit, are lower than Baowaly et al. [2019] and Barbosa et al. [2016], suggesting better prediction accuracy across genres, though direct comparisons are not entirely fair due to different corpora and languages.

5. Final remarks

This paper presents a study on the prediction of review helpfulness in Portuguese, evaluating classification and regression methods. There are two main contributions of this work: the creation, annotation, and availability of a new dataset for the task, with a new domain for Portuguese, and the adaptation and evaluation of a state-of-the-art method for Portuguese.

We also demonstrate that our approach matches results obtained in the literature for other corpora and languages. For the interested reader, the SteamBR dataset and the investigated methods are available at `https://github.com/germanojorge/SteamBR` and at the POeTiSA project web portal (https://sites.google.com/icmc.usp.br/poetisa).

References

Abukausar, MD., S.dhaka, V and Kumar Singh, S. (2013) "Web Crawler: A Review. In: International Journal of Computer Applications, v.63, n.2, p.31-36

Balage Filho, P., Pardo, T. and Aluísio, S. (2013) "An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis. In: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, p.215-219

Barbosa, J., Moura, R. and Santos, R. L. (2016) "Predicting Portuguese Steam Review Helpfulness Using Artificial Neural Networks In: Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web (pp. 287-293).

Bertaglia, T. F. C. (2017) "Normalização textual de conteúdo gerado por usuário". Tese (Mestrado em Ciência da Computação) – Instituto de Ciências Matemáticas e de Computação.

Blei, D., Ng A.y and Jordan, M. (2003) "Latent Dirichlet Allocation", In: Journal of Machine Learning Research., p.993–1022.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. arXiv preprint arXiv:1810.04805.

Friedman, J. (2001) “Greedy function approximation: A gradient boosting machine.” In: The Annals of Statistics, v.20, n.5., p.1189-1232.

Kim, S. M., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006). “Automatically assessing review helpfulness”. In: Proceedings of the 2006 Conference on empirical methods in natural language processing., p. 423-430.

Krishnamoorthy, S. (2015). “Linguistic features for review helpfulness prediction”. In: Expert Systems with Applications, 42(7)., p. 3751-3759.

Le, Q. and Mikolov, T. (2014). “Distributed Representations of Sentences and Documents”. In: Proceedings of the 31st International Conference on Machine Learning, 32(2):1188-1196

Liu, B. (2012). “Sentiment analysis and opinion mining”. In: Synthesis lectures on human language technologies, 5(1), 1-167.

Liu, J., Cao, Y., Lin, C.Y., Huang, Y., Zhou, M. (2007) “Low-quality product review detection in opinion summarization”. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)

Lu, Y., Tsaparas, P., Ntoulas, A., & Polanyi, L. (2010). “Exploiting social context for review quality prediction”. In: Proceedings of the 19th international conference on World wide web., p. 691-700.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). “Distributed representations of words and phrases and their compositionality”. In: Advances in neural information processing systems, 26.

Mudambi, S. M., & Schuff, D. (2010). “Research note: What makes a helpful online review? A study of customer reviews on Amazon.com.”. In: MIS quarterly, 185-200.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). “Linguistic inquiry and word count: LIWC 2001”. In: Mahway: Lawrence Erlbaum Associates, 71.

Sousa, R. F. D., Brum, H. B., & Nunes, M. D. G. V. (2019). “A bunch of helpfulness and sentiment corpora in Brazilian Portuguese”. In: Symposium in Information and Human Language Technology – STIL. SBC

Sousa, R., & Pardo, T. (2022). “Evaluating Content Features and Classification Methods for Helpfulness Prediction of Online Reviews: Establishing a Benchmark for Portuguese”. In: Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis., p. 204-213.

T. A. S., & Aluísio, S. M. (2014, May). “A Large Corpus of Product Reviews in Portuguese: Tackling Out-Of-Vocabulary Words”. In: LREC., p. 3865-3871.

Zhang, Z., & Varadarajan, B. (2006). “Utility scoring of product reviews”. In: Proceedings of the 15th ACM international conference on Information and knowledge management., p. 51-57.