École Polytechnique Fédérale de Lausanne

GameLab

Can the style and wording in critical reviews of video games predict its PEGI labelling?

Machine Learning

Anne Donnet, Haley Owsianko, Mickael Rey

December 17, 2020

I. Introduction

Every year, new games are designed and released in an increasing manner. [1] Even if a majority of video game players are adults, the number of children playing video games is still substantial - 21% of gamers in the USA in 2020 are under 18 years old [2]. To help parents decide if a video game is adapted for their children, the Pan European Game information (PEGI) was founded in 2003. PEGI creates standard age labels for video games, and classifies them with it. In 2017 nearly 30'000 games had been thus labeled [3].

In parallel, with the game industry development, press dedicated to video games gained more and more importance. Video games press websites generate a lot of traffic, and some people use articles from website they trust to choose which game to play. Usually, press articles are meant to highlight the main characteristics of a game, and to give the redaction's opinion. But do the PEGI labels appear through the writing? Or to put it differently, how do the style of writing correlate with the PEGI labels? The answer to this question could help decide whether game reviews can be a good tool for determining the appropriate age to play a game.

II. Methods

Using the PEGI labels as a base for game classification seemed a logical choice, since this norm has been widely used in Europe in the last decades. We decided to use a single website specialized in video game reviews, *jeuxvideo.com*. In the scope of this project, it seemed appropriate to focus on a single website, and jeuxvideo.com had characteristics that suited our goals.

- The reviews on this website are mostly those of well known games, that consequently are classified by PEGI.
- All of the reviews since the creation of the website (1997) are available.
- jeuxvideo.com is based in France and targets a french-speaking audience. It is thus is in the scope of the European PEGI classification.
- The reviews are all written in french, so there was no language-based selection to make on the reviews.
- On a more practical note, the PEGI classification of a game is most of the time directly available on the website, thus simplifying the scrapping.

A. Feature vectors creation

18'349 reviews were extracted from jeuxvideo.com, and the 11'772 that had an age classification were kept. Then, a vocabulary was built in two parts: first, all of the unique words were extracted, and then all the combinations of two consecutive words (bi-grams) were selected. In both cases, stop-words were removed (see list on github repository). Than, an occurrences boundary selection was made, and a stemming was performed. The occurrences boundary selection consisted in discarding all of the unique words that appeared less than 3 times and all the pair of words that appeared less than a hundred times. The first boundary was chosen based on the intuition that a word with one or two occurrences in the whole set of reviews could not be very relevant on the age classification. For the upper boundary, the limit under which the pair of words seemed to have no logical link was selected.The bi-grams vocabulary was further reduced by only accepting those that contained at least one key-word from a list. This list can be found on the repository and was constructed by manually selecting words appearing in the top of bi-grams occurrences that changed meaning by being matched with different words. The corresponding vocabulary size can be seen in Table II-A. The restriction of the vocabulary aimed at reducing the size of the input data in order to improve the training of our models.

TABLE I
Vocabulary word count

	1-word	2-words
Total	100'158	2'350'050
Without stop-words	99'777	3'584'811
Within occurrence boundaries	49'074	2'011
Containing key-words only	∅	400
After stemming	27'206	391

Based on this final vocabulary of 27'597 words, a *term frequency–inverse document frequency* (tfidf) method was used on the collection of reviews and the resulting vectors were used as feature vectors is some models.

A direct mapping from texts to vectors of smaller dimension was also considered. Implementations of such embeddings exist in python [4], so we tried to apply it to our dataset. This produced a vector of dimension 300 for each review, that was used for the predictions.

B. Algorithm and parameters

Once feature vectors were available, different methods were applied to try to predict the best classification.

On the one hand, the feature vectors of small dimension extracted by embedding the reviews were used with the help of a simple k-nearest neighbors algorithm. This model performed best with k = 7. Even though the implementation of text2vec was already available in python, it was interesting to see how this algorithm performed on our dataset.

On the other hand, Tf-idf vectors belonged to a large dimension space. Consequently, dimension reduction techniques were considered.

Considering the fact that this project was focused on NLP, using word embedding seemed a promising solution. To do so, multiple approaches were considered.
First, a pre-trained model of Word2Vec [5] was applied on our dataset. This embedding was trained on the french Wikipedia, which allows the use of a large corpus in multiple contexts. To combine this embedding with our corpus and previously defined vectors, a vector representing each text was created by weighting the embedding of a word by its tf-idf coefficient. Once this transformation was made, classical machine learning methods such as tree based ones (gxboost for instance) were applied. Unfortunately, there were differences between the embedding dictionary and the vocabulary, probably due to video games specific words that are not often used in everyday language. This probably reduced the efficiency of this method, that was already not as good as the one of other methods.

To avoid the previous differences in vocabulary and to learn context specific embeddings, we tried to use only the textual reviews at our disposal to train our models. The use of neural networks allows to implicitly map the words into a smaller space, from which predictions are made by the higher layers. To this end, a classical neural network was defined. This network was composed of 3 fully connected hidden layers, with an input of the vocabulary size and an output size of 5, returning scores for each pegi class. The architecture of this network can be seen figure 1. The input size corresponds to the size of the total vocabulary (100 158 words) since this is the dataset that gave the best results.

We also attempted to turn the question into a regression model, given that smooth transitions between classes are technically possible. However, the networks trained that way ended up having poor results.

This main network was trained for fifty epochs, using a learning rate of 10^{-4}. The training converged for higher values for the learning rate, but we decided to try to make sure we had a proper local minimum. Naturally for a classification problem, cross entropy loss was used. The optimisation was done using pytorch's implementation of the Adam algorithm. The data was separated into batches of 100 data points.

Finally, a different type of neural network was considered. In order to use the texts as so and not just unordered bag of words, recurrent neural networks were tested. More specifically, lstm (Long Short Term Memory) networks were used on our data to try and predict the pegi classification of a game based on its review. This kind of neural networks uses previous output as a complementary input, and hence the final output is based on the entirety (to a certain extent) of an ordered input, in our case the words of a review. The considered network had an embedding layer, followed by lstm layers. Unfortunately, we did not manage to make this network converge. Maybe this is due to a too small number of available training datapoints, or to insufficient adjustments of the hyper-parameters of this network.

III. Results

The model that ended up performing the best was a naïve neural network, shown in figure 1, with an accuracy of around 69.5%, reaching up to around 81% if we switched to a classification based on label colour—green for +3 and +7, orange for +12 and +16, and red for +18. Given the number of classes, this is far better than random, but still a long way from a good classifier.

Given the number of weights in the model, it is surprising that a relatively small training set was enough to reach such a good accuracy.

With a text-embedded input, the models considered ended up maxing out around 66% classification accuracy. Hence, we were unable to find a model that performed better than the naïve neural network. Colour-based classification was also lower, though it ended up around 79%, which is no that far from the neural network's accuracy. Assuming that our attempted neural network architectures were reasonable, this could mean that using word embedding removes information about the specific text and thus makes it more difficult to recognise the exact rating of a game, but can become almost as good as the giant network at interpreting the general tone of a review.

A speculation for an explanation for this would be that two words that are semantically similar might not

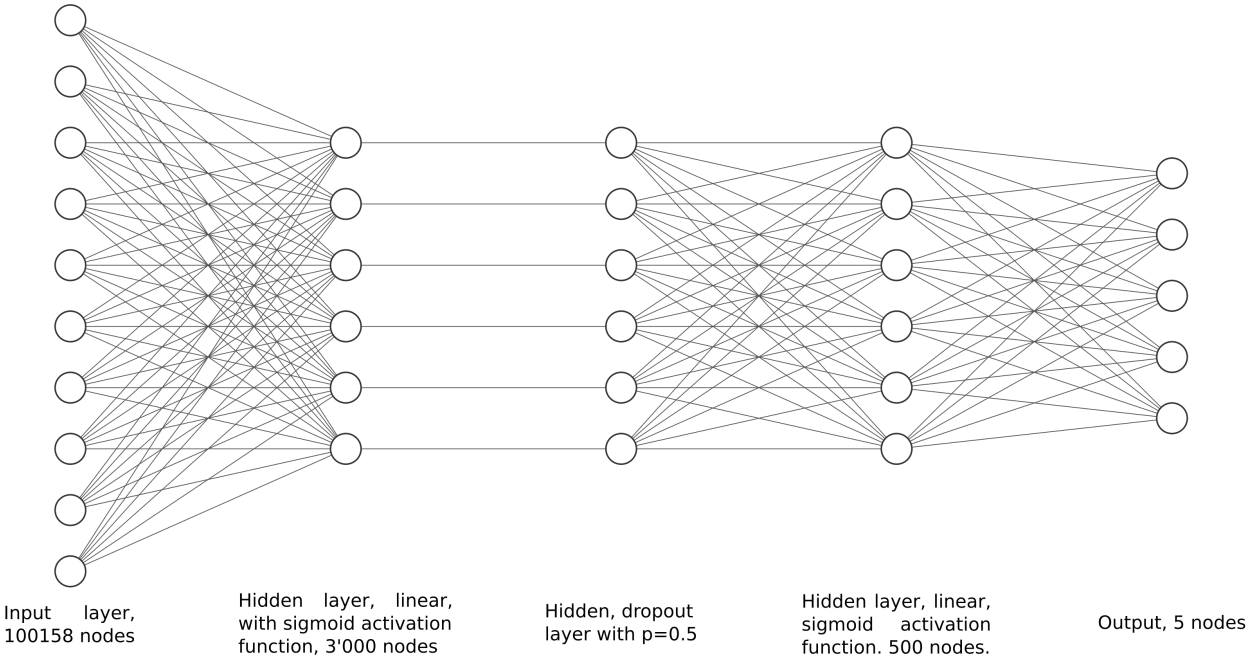
Input layer, 100158 nodes
Hidden layer, linear, with sigmoid activation function, 3'000 nodes
Hidden, dropout layer with p=0.5
Hidden layer, linear, sigmoid activation function. 500 nodes.
Output, 5 nodes

Fig. 1. Architecture of the neural network

Method	Pegi accuracy	Color accuracy
tf-idf + word embedding	0.60	0.73
text embedding	0.66	0.79
Fully connected neural network	0.70	0.81

TABLE II
PREDICTION RESULTS BY DIFFERENT METHODS

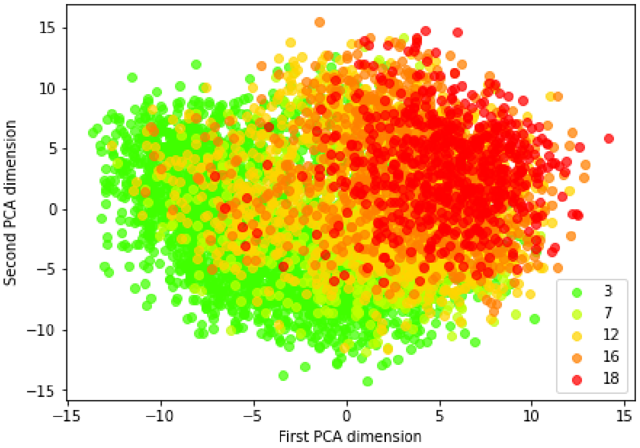

Fig. 2. Embedding of the reviews projected on a 2-dimensional space.

be used in the same contexts. For instance, when in a general video game context, ”eliminate”(”éliminer”) and ”kill”(”tuer”) usually mean the same thing when referring to enemies, but a reviewer might be more inclined to use ”eliminate” when writing a review a game with a lower age rating.

The embedding allowed nonetheless to visualize the texts on a small space, by performing a Principal Component Analysis (PCA) on the extracted vectors. Figure 2 clearly shows a tendency in the classification according to the first 2 dimensions of the projection of the vectors computed with text2vec. Regardless, our methods using these vectors did not reach higher than around 66% of accuracy.

We deemed it interesting to look at a few examples of misclassified reviews. First up, we have the more severe category, games misclassified downward, i.e. games that are rated 18+ but that our algorithm classified as 3+. We took a look at three reviews that were misclassified like this.

First up, ”Fallout 4: Vault-Tec Workshop”. The reason for the wrong label is quite obvious here. This DLC—Downloadable content; an add-on to the game—revolves around being given extra options to build settlements. Naturally, much crafting related terminology appears in the review, overshadowing the few occurrences of words that a human would associate with a higher age rating. While a human reading the review would definitely understand it isn't about a family friendly game, it is easy to see why a neural network might misinterpret it.

Second, "Driver 76". This example is even more understandable than the previous one, as the review spends little time discussing the more mature themes of the game. The only part of the review that can indicate the nature of the game is the part where the author mentions the game's storyline, which involves gangs. However, no brutality is described in that part of the review. Even a human might end up underestimating the game's age rating, although most likely not quite as extremely as the neural network, given the mentions of gangs.

Last, "OneChanbara : Bikini Samurai Squad". It is hard to understand why the neural network misclassified this review so badly, as the text is full of sexual and violent terminology. A human would not hesitate at all before realising that this is not a family friendly game, and it is also difficult to imagine why a neural would have trouble with it, given the frequency of those terms.

IV. Discussion

In the end, no features pre-processing really improved our results. This could be due to the size of our dataset that may be not big enough to gain anything by being pre-processed. Also our results may be improved by augmenting the data set.

Nonetheless, it was possible to predict the PEGI classification of game based on its textual review with a satisfactory accuracy. By being clearly better than a random choice, the precision we achieved shows that a link does exists between the vocabulary used in reviews and the age rating of the game. Having even better results when classifying the PEGI colors shows that a non-negligable part of our misclassified text are classified in an age category not far from theirs. This could be used as an argument to ascertain that journalistic game reviews can be used by parents to know if a game is adapted for their child's age category.

Furthermore, those results could be used as a starting point for an automated analysis of video game speeches, a new tool for this really active search domain[6], that potentially has a good amount of exploitable data [7].

References

[1] Number of games release on steam. https://steamspy.com/year/.

[2] Age of video games players in 2020 in the us. https://www.statista.com/statistics/189582/age-of-us-video-game-players-since-2010/.

[3] Statistics about pegi classification. https://pegi.info/page/statistics-about-pegi.

[4] Dmitriy Selivanov *et. al.* text2vec package. http://text2vec.org/.

[5] Greg Corrado Tomas Mikolov, Kai Chen and Jeffrey Dean. Efficient estimation of word representations in vector space (word2vec). *Google Inc.*, 2013.

[6] Boris Krywicki. La présence du comportement investigateur dans la presse jeu vidéo – positionnement dans le champ, typologie et analyse. Thesis paper, 2016.

[7] Yves Breem Boris Krywicki. *40 ans de magazines de jeux vidéo en France*. Presse Start, 2020.