ISSN 2616-7433 Vol. 2, Issue 12: 70-80, DOI: 10.25236/FSST.2020.021212

Prediction of the Most Popular Game Tags on Steam under the Influence of Covid-19 Based on Machine Learning and Natural Language Processing

Jihao Zhang[1], Hongzhi Zhao[2], Zhicheng Chen[3], Zihan Song[4]

1 Department of Foreign Language, Huazhong University of Science and Technology, Wuhan 430074, China
2 Department of Computer Science and Technology, North China Electric Power University, Beijing 102206, China
3 Department of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China
4 Department of Mathematics and Statistics, South Central University for Nationalities, Wuhan 430074, China

ABSTRACT. *Considered the influence of COVID-19 epidemic on video game markets, this paper proposes using Python to find the best model of predicting the most popular game tags on Steam game platform. Four of the models of predicting the best-selling games on Steam are based on Lasso linear regression, support vector machine, decision tree and random forest. Then people can predict the most popular tags by sum the tags of these games. Another one of the models of predicting the most popular game tags directly are based on natural language processing and random forest. This research succeeds to provide a game tags predicting model combining the catastrophic events and machine learning for the first time.*

KEYWORDS: *Machine learning, Linear regression, Svm, Natural language processing*

1. Introduction

Since January of 2020, many people were mandated to stay at home to avoid contacting the COVID-19 virus. This epidemic is a disaster to humankind exerting influence on many aspects including video game market. Referring to a popular article called What is happening with video game sales during coronavirus 1, people have increased interest for online video games because of the outbreak of the epidemic and the isolation at home leading to an increase of online gaming sales. As a result, the availability of video games and purchase behaviors during COVID-19 epidemic make for a viable study.

Catastrophic events greatly affect human behaviors, so the purchase behaviors will become a valuable study during the epidemic period. Cantin et al. (2016) present some mathematical results concerning the PCR system (Panic-Control-Reflex), which is a model for human behaviors during catastrophic events.[13] This model has been proposed to better understand and predict human reactions of individuals facing a catastrophe, in a context of an established increase of natural and industrial disasters. In a teaching period of more than one and a half months, nearly 270 million colleges, high school, middle school, and elementary school students in China have conducted normal course studies online (Zhou et al. 2020[12]). Before that, Cantin et al. (2016) have presented some mathematical results concerning the PCR system (Panic-Control-Reflex), which is a model for human behaviors during catastrophic events.[13] Verdière et al (2014) have proposed a SIR (Susceptible Infected Recovered Model)-based mathematical model taking into account the psychological reactions of the population in situations of disasters, and studied their propagation mode.[14]

In the field of game sales prediction, there have been a number of research studies conducted. Ahn et al. (2017) examined Steam (one of the biggest and most popular video games platforms) using the Bass diffusion model to identify the factors of success for games that can help drive industry growth in the future.[15] Marcoux and Selouani (2009) advance a new approach based on connectionist and subspace decomposition methods.[5] A tool is designed to support company management in the process of determining expected sales figures. To forecast the trend of new games coming out every year, Cheuque et al. (2009) test the potential of state-of-the-art recommender models based on Factorization Machines (FM), deep neural networks (DeepNN) and one derived from the mixture of both

ISSN 2616-7433 Vol. 2, Issue 12: 70-80, DOI: 10.25236/FSST.2020.021212

(DeepFM).[17] They also analyzed the effect of the sentiment extracted directly from game reviews. Lin et al. (2019) conducted a preliminary study to understand the number of game reviews, the complexity to read through them, studied the relation between several game-specific characteristics and the fluctuations of the number of reviews that are received on a daily basis.[18] Bais et al. (2017) harness Support Vector Machines (SVM), Logistic Regression, Multinomial Naive Bayes, Turney's unsupervised phrase-labeling algorithm, and a lexicon-based baseline to build a binary sentiment classifier.[19]

All the studies about game data analysis and prediction provide useful methods that can be utilized all the time. However, the changes in people's demand for video games under the influence of a catastrophic event such as the COVID-19 epidemic are not considered. As a result, this article focuses on the changes in people's demand for different types of video games under the influence of the COVID-19 epidemic. The main question that this paper explore is to perform staged game data analysis during the epidemic to find out the characteristics of popular games of different tags, and finally predict the next stage of the most popular game tags. In addition, the predicting model could also help improving the precision of gaming recommendations to players on Steam and may help bringing higher profits for the platform.

This paper uses four basic methods of machine learning including linear regression, support vector machine, decision trees and random forest to build a model of predicting the weekly selling amount of games on Steam to show the popularity changing and take advantage of some algorithms to reduce the deviations. In addition, this research makes use of natural language processing methods to deal with the most recent game reviews so that a precise prediction of the increase of sales for certain games per week and the most popular tags in the next stage could be provided.

2. Experimental section

2.1 Data

A.Sales

This research make use of the website *Steamspy* as our source of Steam games data. This paper choose the top 100 games of the best sales from 1 January 2020 to 31 March 2020 on Steam, and crawled everyday worth of data for each of the games mentioned before from 1 January 2020 to 28 April 2020 so that the data could be split into 17 weeks. This dataset of the whole 17 weeks contains 49255 pieces of data altogether. Tab. 1 shows the features used in this dataset.

Table 1 The features and description used for the Steamspy data

Item	Description
steam_id	The unique identity of each video game on Steam platform.
owners_before	The owners of each video game at the start of the selected period.
price	The sale price of each video game
max_discount	The max discount percentage of each video game in the selected period.
sales	The sales amount of each video game in the selected period.

B.Review

This paper leverages a module called "steamreviews" from PyPI(The Python Package Index) downloading 458406 reviews of the top 100 games mentioned above in April.

C.COVID-19

A dataset called *Novel Corona Virus Dataset (COVID-19)* created by *Anjana Tiha*, from *Kaggle* is used in the research. By extracting the everyday confirmed cases of COVID-19 in every country, the weekly global confirmed

ISSN 2616-7433 Vol. 2, Issue 12: 70-80, DOI: 10.25236/FSST.2020.021212

cases are summed up so that the data of coronavirus epidemic from 1 Jan. 2020 to 26 Apr could be used as a feature for the prediction.

D.Tags

In this research all the game tag information until 26th April 2020 on Steam from *Steamspy* website is downloaded as a dataset containing 339 tags altogether. Tab. 2 shows the features used in this dataset.

Table 2 This table shows the features and description of tags used for the Steamspy data

Feature	Definition
name	The spelling of the tag
games with this tag	The number of the video games which are attached by the tag
Votes for this tag	The total number of players who vote for the tag
userscore (median)	The median score given by the players of the video games attached by the tag
price (median)	The median price of video game attached by the tag
playtime (median)	The median daily playtime of each video game attached by the tag
Owners (median)	The median amount of owners of each video game attached by the tag

1.2 Methodology

In this research, several basic methods of machine learning are used to manage our data. The code is running on Jupyter Notebook with Python 3.7.

A.Managing the Weekly Information

Steps

1) Split Epidemic and the Data into 17 Weeks

2) Selected 6 Features as Our Variables Make It into a Vector for the Prediction. Tab. 3 Shows the Features Used for the Prediction Model.

Table 3 The features and description used for the prediction model

Feature	Description
week	The number representing the *n*-th week.
steam_id	The unique identity of the video game on Steam platform.
owners_before	The owners of the video game at the start of the week.
price	The sale price of the video game
max_discount	The max discount percentage of the video game in the selected period.
covid_confirmed	The number of confirmed cases of COVID-19 patients of the week

3) Visualize the Relation between Sales and Every Feature and Calculate the Standard Difference of Each Feature. Figure 1 Shows the Visualization Mentioned Above.

ISSN 2616-7433 Vol. 2, Issue 12: 70-80, DOI: 10.25236/FSST.2020.021212

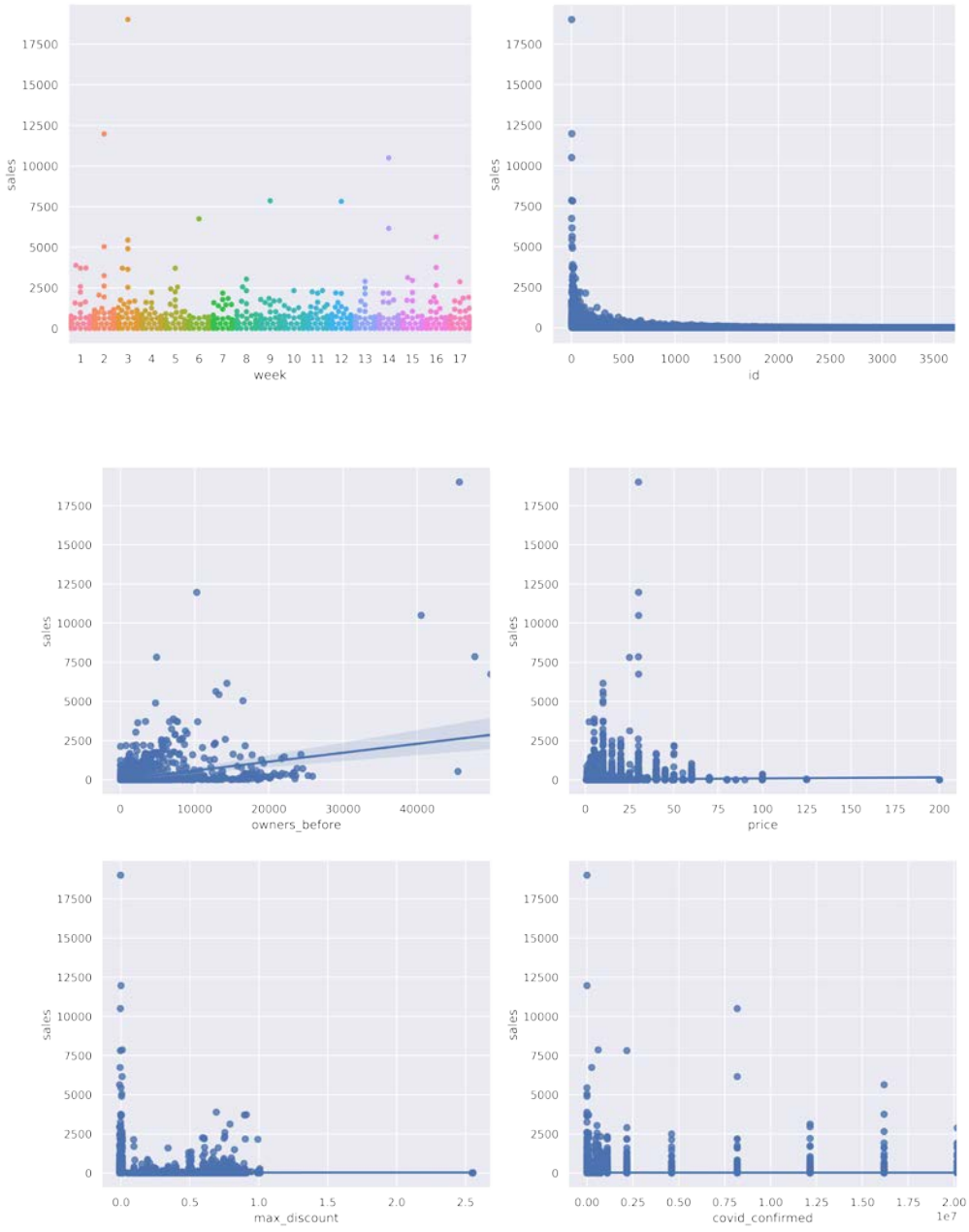

Fig.1 The visualization between sales and every feature of the data.

4) Pick Up the Data from January to March as Our Training Set and the Data in April as Our Test Set Then Shuffle These Data Randomly. Later Train the Dataset with Some Basic Methods of Machine Learning and Build Several Models.

5) Test the model with the test set. The tests take the sum of the square of the difference between the predicted value and the true value and use them build the loss function and then moderate the parameters. The final target of our model of linear regression is to give more accurate prediction of the difference of the owners amount per week by providing weekly information of one game on Steam. Then summarize the tags of the top 10 most popular games by judging the increasing amount of the owners and find the most popular tags

A lgorithms

ISSN 2616-7433 Vol. 2, Issue 12: 70-80, DOI: 10.25236/FSST.2020.021212

This research leverages the following methods to build the game tag predicting model.

1) Lasso Linear Regression

Linear regression is a popular algorithm which is used to predict the value of a variable based on the value of other variables. Lasso is a good advanced linear regression algorithm which could build a model for estimating sparse coefficients. Since our variables are not too many and the difference of the amount and weight is relatively big, this paper makes use of the Lasso to avoid overfitting.

2) Support Vector Machine

Support Vector Machine(SVM) is a good algorithm for classifying when data is not so much. It is also efficient to the high dimension feature vector. As a result, this research also take advantage of SVM to build a model for the prediction.

3) Decision Trees

Decision Trees is also an ideal choice since it does not require vast amounts of data and typically yields a precise model for regression predictions. As a result, Decision Trees tends to be a good algorithm to help this researcg classify and make predictions.

4) Random Forest

Random Forest algorithm is a great method to extract import features. The difference of the weight of our features tends to be large, so this method is also proper for the model-training.

B.Managing the Reviews

To better predict the most popular tags, this research select the reviews including the words of or about the tags and then study them separately.

Steps:

1) Using the Python NLTK package, this paper performed preprocessing steps to clean the data, such as removing stop words and extracting stems from the comments of each label. The code loop through 40% of the comments and all tags to match them with the test data, and then leveraged the Random Forest algorithm to classify the tags of remaining 60% of the comments. Tab.4 shows the model of the dataset for random forest classifier

Table 4 The model of the dataset for random forest classifier

	body	body_stem	body_tokenize	tag
0	bought early access game	bought earli access game	[bought, early, access, game]	early access
1	one best vr games	one best vr game	[one, best, vr, games]	vr
2	great arpg servers	great arpg server	[great, arpg, servers]	rpg
3	really fun platformer	realli fun platform	[really, fun, platformer]	platformer
4	tested positive covid	test posit covid	[tested, positive, covid]	covid
...	...	...	...	...
11705	great survival online	great surviv onlin	[great, survival, online]	survival
11706	finaly good arpg game	finali good arpg game	[finaly, good, arpg, game]	rpg
11707	fun survival feel	fun surviv feel	[fun, survival, feel]	survival
11708	great game like rogue likes	great game like rogu like	[great, game, like, rogue, likes]	rogue like
11709	really good mmorpg trying	realli good mmorpg tri	[really, good, mmorpg, trying]	rpg

11710 rows × 4 columns

2) In addition, this paper divides the comments of different tags every day. Similarly, the number of COVID-19 confirmed cases are calculated every day, and linear regression is used to train a model to show the relationship between user's preference for the game and the severity of the epidemic. Next, this research uses the Cost Function and Gradient Descent method to test and optimize the results.

ISSN 2616-7433 Vol. 2, Issue 12: 70-80, DOI: 10.25236/FSST.2020.021212

3) Finally, Our Group Use Word Clouds and Some Other Visual Methods to Show Players' Interest in Different Game Tags At Different Times.

2. Results and discussion

After training, four models built by four algorithms which are Lasso Linear Regression, Support Vector Machine, Decision Trees and Random Forest are set up. After testing them separately with testing dataset and Compared with the actual value, the visualization of the result of every model comparison charts are performed.

Firstly, this paper builds a model by using linear regression. This research chooses Lasso Linear Regression to avoid over-fitting. After several times of trying, λ=5 (λ means the learning rate of the linear regression) is found to be the best value. After training, our group test the model with the training set and testing set. The fitting score of training set is 0.2017 and the fitting score of testing set is -0.018. However, the best fitting only reaches 1%, which tends to be relatively low. This calculate the standard difference between each real value and predicting value of the training set and testing set. The standard difference between real value and predicting value is 15569.64. After the analysis, the problem is found to be that the features are not enough for the linear fitting. Fig. 2 is the visualization of the prediction results of lass linear regression model.

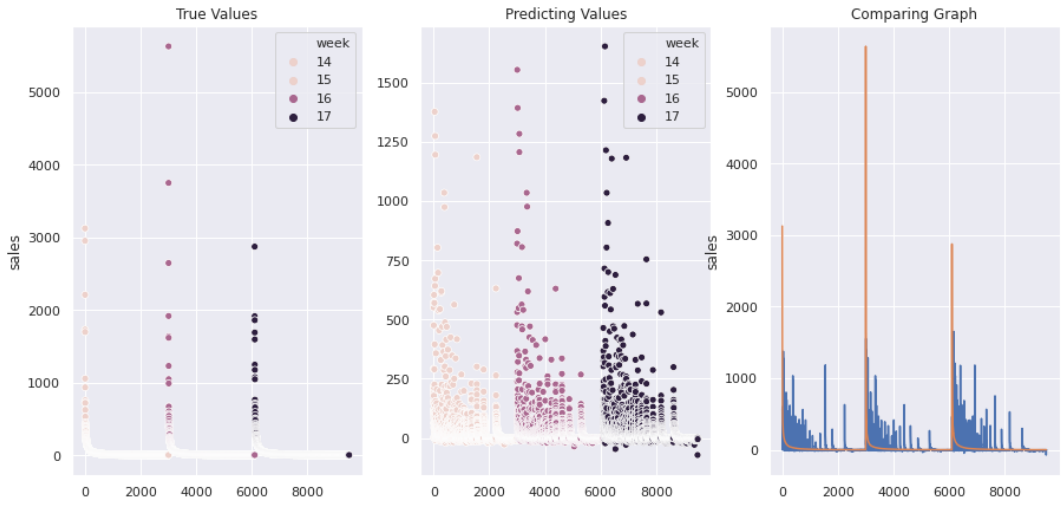

Fig.2 The visualization of the prediction results of lasso linear regression model.

Then, our group build a model with support vector machine(SVM). Fig. 3 shows the relation between the prediction results of SVM model. In this research the fitting effort tends to be best when setting c=1.5 (c means the value of punishment) and setting the kernel function as Gauss Function. After training, the model is also tested with the training set and testing set. The fitting score of training set and testing set are both 0.25, which is much better than linear regression. However, this result is also not high. The standard difference between the real value and the predicting value is 15172.95, which is still such a high amount. After the analysis, the reason tends to be that the features are too few to make a reasonable classification.

The fitting score of testing set is 0.26, but the fitting score of testing set reaches 1.0. Therefore, there must be an over-fitting of the model. However, no matter how to moderate the parameter, the over-fitting situation is hard to be avoided. As a result, the conclusion is that our data is too little to build an ideal prediction model by using decision tree. This paper also calculates the standard difference between real value and predicting value, which is 14256.440623625807. This is also a relatively high amount. Fig. 4 shows the visualization of the predicting results of decision tree.

ISSN 2616-7433 Vol. 2, Issue 12: 70-80, DOI: 10.25236/FSST.2020.021212

Fig.3 The visualization of the prediction results of SVM model.

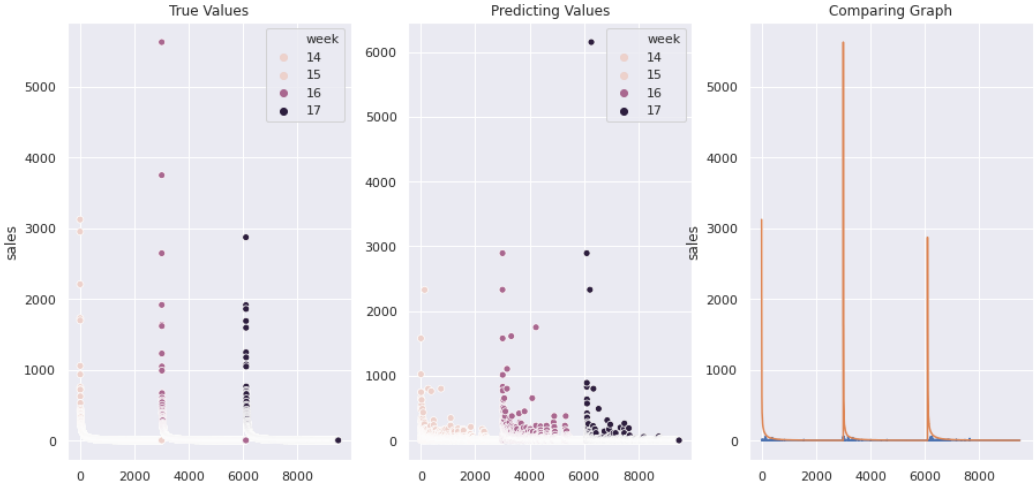

Fig.4 The visualization of the prediction results of decision tree model.

Finally, random forest algorithm becomes the last selection in this research to build a model to make predictions. The random forest is good to select feature of the most weight. After training and setting the estimator depth as 1000 and the random state as 5, the best model of this algorithm is performed. Then the model is also tested with both the training set and the testing set. The fitting score of testing set is 0.59, and the fitting score of testing set reaches 0.92. This tends to be the best result and model up till now. However, 60% is not a precise prediction rate yet. The standard difference between the real value and the predicting value still reaches 14256.44. From the analysis, the reason is still mainly from that the features are not enough to raise the precising rate. Fig. 5 shows the visualization of the predicting results of random forest.

ISSN 2616-7433 Vol. 2, Issue 12: 70-80, DOI: 10.25236/FSST.2020.021212

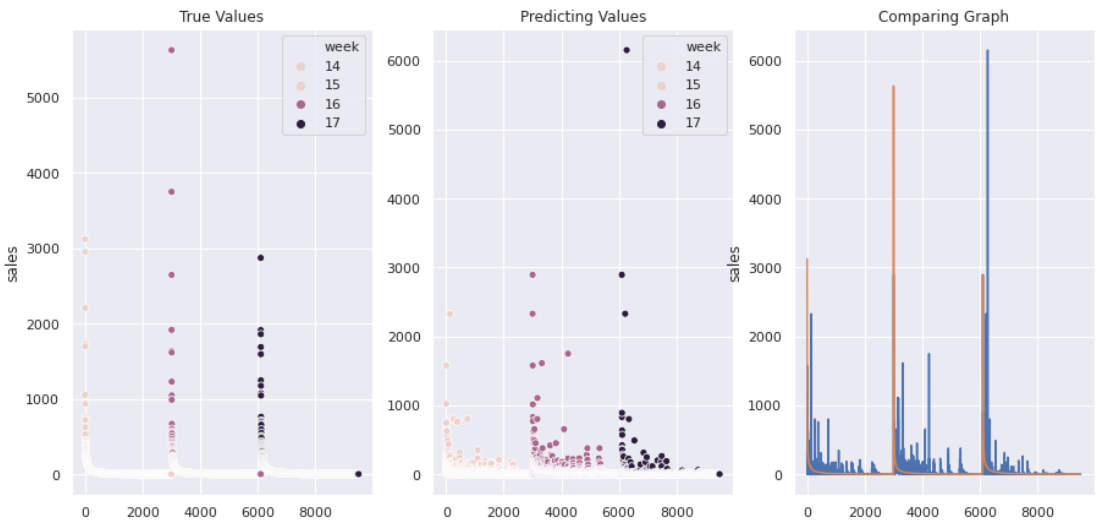

Fig.5 The visualization of the prediction results of random forest model.

As a result, reviews become the other vital path to find the most popular tags.

In fact, because most reviews have few words, many are meaningless, resulting in few useful reviews. In addition, there are too many types of game tags, which makes us unable to obtain satisfactory results when using the random forest algorithm to classify the data, so only the tags being concerned with, combined with the random forest algorithm and some simple methods of extracting tags are selected to classify 450,000 reviews.

When leveraging random forest algorithm, the code performs the best parameters of train_test_split, which are: test_size = 0.8, random_state=1, and the precision F-score reached (0.7535036296414998, 0.6498128150878392, 0.6797953894251205, None), using the GridSearchCV, the best parameters of the randomforest classifier (max_depth: 50, n_estimators: 500, random_state: 2000) are got.

The last result is relatively clear. This research uses Linear Regression and simple coordinate relationships. Fig. 6 performs that as the epidemic becomes increasingly serious, players' interest in the Adventure games at first increased, perhaps because of the quarantine at home, however when the epidemic became increasingly serious, the interest in Adventure games decreased.

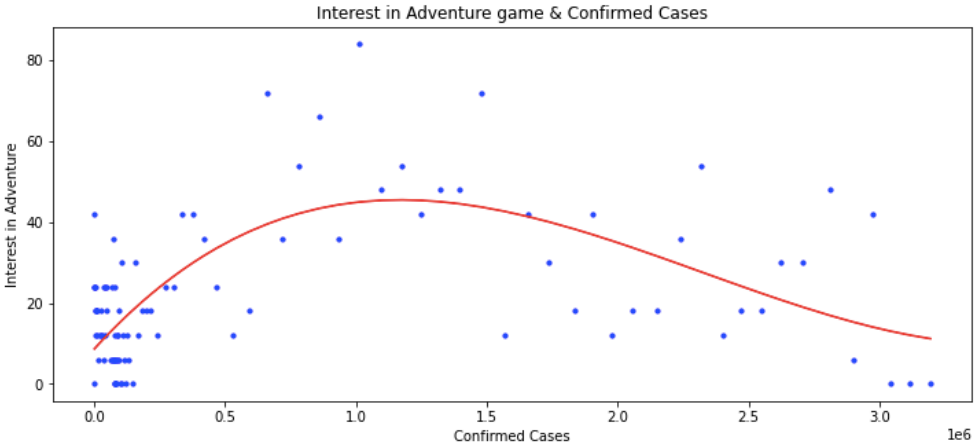

Fig.6 The change of the interest in adventure games when the epidemic becomes serious

ISSN 2616-7433 Vol. 2, Issue 12: 70-80, DOI: 10.25236/FSST.2020.021212

To compare different tags in a direct way, this paper put the curves of different tags in one chart. In Fig. 7, different tags actually change in the same way, there's little evidence that players have some preference for some tags during the epidemic.

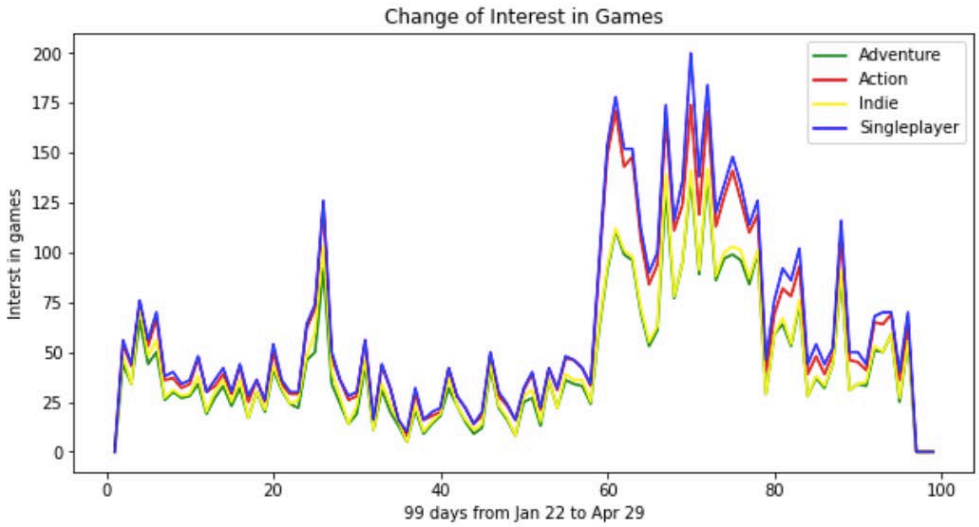

Fig.7 The change of players' interest in different tags from Jan 22 to Apr 29

In addition, Fig. 8 shows that as time goes by, players are paying more and more attention to the epidemic. The mention rate of words such as 'coronavirus', 'quarantine', 'epidemic' is increasing at first, and goes down with time going by.

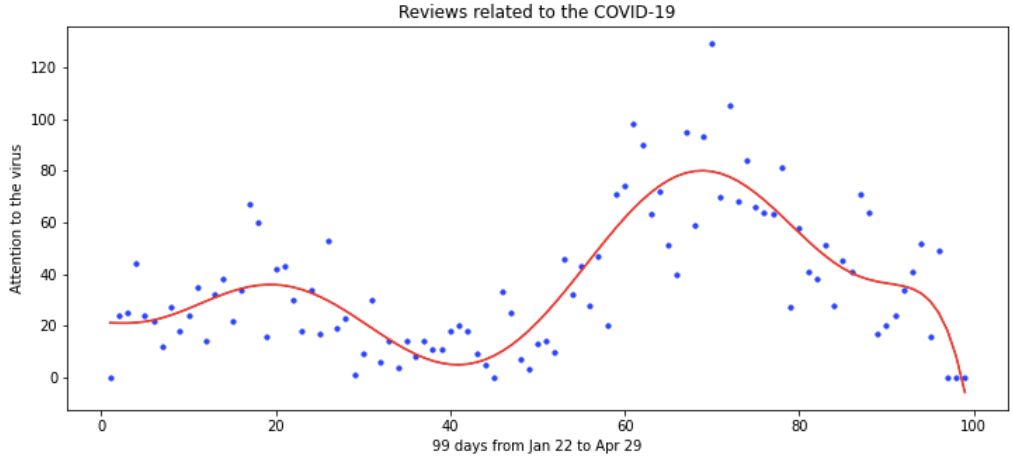

Fig.8 The change of the mention rate of words related to the covid-19 epidemic

Finally, this research uses the word cloud to show the most popular game tags in these three months, which is shown in Fig. 9.

ISSN 2616-7433 Vol. 2, Issue 12: 70-80, DOI: 10.25236/FSST.2020.021212

Fig. 9 The word cloud of the most discussed tags of the games from Jan to Apr.

4. Conclusion

Compared with the two results of predicting model and natural language processing, the result of the prediction of the top5 most popular tags of video games on Steam in the first week in May 2020 are: advanture, fps, rpg, vr and early access. This result is precise after the examination after a week.

This research combines the video game platform and the situation of the catastrophic events for the first time and succeeds to make a precise prediction of the most popular game tags. This is a meaningful result because when the change of purchasing behaviors under the influence of the catastrophic events are considered and the recommendation could change correspondingly, it will be easier for the video game companies and the game platforms to grasp the players. As a result, combined with the predicting model built in this research better recommending systems of the game platforms and the video game company online stores could be designed more effectively so that more profits will be made, which tends to be beneficial for the development of the video game market and even the whole market of the virtual products.

Acknowledgment

First and foremost, we would like to show my deepest gratitude to my professor, Dr. Patrick Houlihan who has provided us with valuable guidance in every stage of the writing of this thesis. We shall extend my thanks to Mr. Huang for all his patience and help. Finally, we would like to thank all my friends for their encouragement and support.

References

[1] Miloš Milošević, Nenad Živić, Igor Andjelkovića (2017). Early churn prediction with personalized targeting in mobile social games, Expert Systems with Applications, no.83, pp.326-332.

[2] Suchacka, Grażyna, Sławomir Stemplewski (2017). Application of neural network to predict purchases in online store, Information Systems Architecture and Technology: Proceedings of 37th International Conference on Information Systems Architecture and Technology–ISAT 2016–Part IV, pp.221-231.

[3] Borbora, Zoheb H, Jaideep Srivastava (2012). User behavior modelling approach for churn prediction in online games, IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Confernece on Social Computing, Amsterdam, Netherland, pp.51-60

[4] Croda, Rosa María Cantón, Damián Emilio Gibaja Romero (2019). Sales prediction through neural networks for a small dataset, IJIMAI, vol.5, no.4, pp.35-41.

[5] Marcoux, Julie, and Sid-Ahmed Selouani (2009). A hybrid subspace-connectionist data mining approach for sales forecasting in the video game industry, IEEE WRI World Congress on Computer Science and Information Engineering, the United States, no.5, pp.666-670.

[6] Thomopoulos, Nick T (2015). Demand forecasting for inventory control, Demand Forecasting for Inventory Control, pp.1-10.

ISSN 2616-7433 Vol. 2, Issue 12: 70-80, DOI: 10.25236/FSST.2020.021212

[7] Dannenberg, and Jeff (2010). Internet based prediction market, U.S. Patent Application 12/331,782, filed June 10.
[8] H. Rodriguez, V. Puig, J. J. Flores, and R. Lopez (2016). Combined holt-winters and GA trained ANN approach for sensor validation and reconstruction: Application to water demand flowmeters, 3rd Conference on Control and Fault-Tolerant Systems, Barcelona, Spain, pp.202-207.
[9] Ioannis Krasonikolakis, Adam Vrechopoulos, and Athanasia Pouloudi (2014). Store selection criteria and sales prediction in virtual worlds, Information & Management, vol.51, no.6, pp.641-652.
[10] Geva, Tomer, Gal Oestreicher-Singer, Niv Efron, et al (2015). Using forum and search data for sales prediction of high-involvement products, MIS Quarterly, Forthcoming.
[11] Rephael Sweary, Michael Eden, and Yaron Golan, Multi-Stage Future Events Outcome Prediction Game, US Patent App. 12/223, 612, 2009
[12] Zhou, Longjun, Shanshan Wu, Ming Zhou (2020). 'School's Out, But Class' On', The Largest Online Education in the World Today: Taking China's Practical Exploration During The COVID-19 Epidemic Prevention and Control As an Example, But Class' On', The Largest Online Education in the World Today: Taking China's Practical Exploration During The COVID-19 Epidemic Prevention and Control As an Example.
[13] Cantin, Guillaume, Nathalie Verdiére, Valentina Lanza, et al (2016). Mathematical modeling of human behaviors during catastrophic events: stability and bifurcations, International Journal of Bifurcation and Chaos, vol.26, no.10, pp.1630025
[14] Verdière, Nathalie, Valentina Lanza, Rodolphe Charrier, et al (2014). Mathematical modeling of human behaviors during catastrophic events.
[15] Ahn, Sangho, Juyoung Kang, Sangun Park (2017). WHAT MAKES THE DIFFERENCE BETWEEN POPULAR GAMES AND UNPOPULAR GAMES? ANALYSIS OF ONLINE GAME REVIEWS FROM STEAM PLATFORM USING WORD2VEC AND BASS MODEL.
[16] Cheuque, Germán, José Guzmán, Denis Parra (2019).Recommender systems for Online video game platforms: The case of STEAM, In Companion Proceedings of The 2019 World Wide Web Conference, San Francisco, the United States, pp.763-771.
[17] Lin, Dayi, Cor-Paul Bezemer, Ying Zou, et al (2009). Hassan, An empirical study of game reviews on the Steam platform, Empirical Software Engineering, vol.24, no.1, pp.170-207.