Research on Virtual Ethnography of Online Games Based on NLP Big Text Statistics and Artificial Intelligence Analysis

Introduction

When it comes to video games, the first thing that comes to mind is that many parents and teachers see them as a predator and want to get rid of them as soon as possible. This reflects the negative image of video games in the eyes of some parents and teachers, who may be concerned that gaming can have a negative impact on their children's learning, physical and mental health, and social resilience.

In fact, the opposition of parents and teachers to the game is also inseparable from the existing test-oriented education system. In the test-oriented education system, parents and teachers regard students' scores as the actual goal, and naturally regard everything that is not conducive to the achievement of this goal as an enemy. In the learning process, if students are not learning spontaneously, they are learning under external pressure. Students who are in a state of "forced learning" will be eager to find novelty and a sense of accomplishment in the game. During this period, teenagers have not yet formed their worldview and values, and they lack resistance to video games that are "designed to make people addicted". Coupled with the psychological impact of the rebellious period of adolescence, video games seem to have become the fuse of parent-child conflicts in many families. There are also many parents who hate video games.

As a result, whether there is a problem with video games per se has become a rather controversial issue. How the game industry should develop and what impact it has on society is also of considerable research value.

This is a complex issue involving psychology, education, economics, philosophy, and even artificial intelligence technology. The world of science and technology is also changing rapidly, and social realities are changing rapidly. How to properly treat video games has become an important issue for mankind in the next few centuries.

Based on this background, we have a strong interest in trying to contribute to solving this puzzle with new perspectives and technologies.

summary

Based on the analysis of macro data, this paper further conducts virtual ethnography of video games through natural language processing technology (NLP) and artificial intelligence analysis methods to conduct large text statistics. Specifically,

it is to efficiently and quickly process the information of a large number of comments on video games on the Internet, so as to objectively, truthfully and comprehensively reflect the views of all walks of life on the Internet on the highly controversial thing of "video games", and draw our conclusions based on macro data.

In the first part, the paper will focus on analyzing online reviews to illustrate the current state of the video game industry in terms of public opinion. Through an in-depth analysis of online reviews, it is possible to understand the public's attitudes and perceptions towards virtual online games, so as to better understand the society's awareness and acceptance of the industry.

Through crawler technology, we collected tens of thousands of comments about video games from four platforms (Weibo, News, Zhihu, Bilibili). The main purpose of this is to fully and comprehensively reflect the views of the "virtual nation" formed by all sectors of society on the Internet platform on the matter of "video games".

These reviews are long and short, rich in content and varied. The evaluation subjects are also highly diversified, from the perspective of industry, there are news media, online bloggers, educators, game industry workers, etc.; In terms of age, there are teenagers, young adults, middle-aged, and so on. And when selecting reviews, it is also based on a random sampling algorithm to ensure that there is no subjective error caused by human preference.

After learning and applying the model, we came to the conclusion that 59.4% of the reviews reflected a positive attitude towards video games.

In the second part, we focused on the analysis of high-information value words and thinking angles in the review. The information value of words is comprehensively reflected by TF-IDF values, and the characteristics of comments on different platforms are vividly displayed in the word cloud. The analysis of the thinking perspective fully reflects the power of natural language processing technology. The results of this analysis have led to some novel and interesting conclusions. This type of analysis was generally in the form of fuzzy estimates and rough estimates before this study, but it has a convincing data base in this study, which has higher decision-making value.

Overall, this paper aims to comprehensively explore the impact of video games and public opinion through the analysis of economic data and online commentary, and provide data and theoretical support for relevant decision-making.

directory

1. LSTM model

1.1 Strategy connotation

In the previous chapter, we mainly understood the role of the game industry in social development from a macro perspective. We analyze the positive social impact of the game industry mainly in terms of market size, cultural export, employment provision, and promotion of technological development. These aspects are all from the perspective of the country to look at the development of the game industry. In this chapter, we prefer to support our argument with more micro-biased data. We try to analyze their emotional attitudes towards games and the development of the game industry from the perspective of the general public. This is the main idea of this chapter.

In this chapter, we focus on the perspective of public opinion to capture, collate and analyze the voices of the public. So we thought about getting discussions and reviews about the development of games and the game industry from a number of websites. Using these comments, we can analyze the user's attitude, emotion, and use artificial intelligence and natural language processing methods to learn from them, and finally apply and predict them.

In this part, we hope that there are good reasons to get reviews from the Internet to reflect the public opinion on the game. First of all, the use of crawler technology to obtain discussion content has the advantage of volume. Because the content that can be crawled is ready-made and not constrained by time and space, in fact, it must be theoretically possible to obtain enough data. Only a large amount of data can reflect the public's tendencies. So this is logically self-consistent. Secondly, machine learning, deep learning, and natural language processing technologies are relatively well-developed, which lays a feasible theoretical foundation for analyzing the sentiment of comments and discussions.

Therefore, it is reasonable to obtain comments through crawlers and analyze the emotional attitudes of a large number of comments to reflect the tendency of public opinion towards the development of the game. It is both feasible and appropriate.

1.2 Data Mining

The first thing you need to do is determine what kind of reviews to get for the final analysis. It is important to note that the acquisition of comments here does not refer to obtaining a dataset for machine learning training, nor is it a test set, but a dataset that ultimately applies the machine learning model for conclusion analysis. We will finally analyze the sentiment of each comment in this dataset, and finally count the proportion, and get the popular emotional attitude.

There are reviews of the game on many websites, such as Bilibili, Weibo, Zhihu, etc., as well as many game-related news. These are all data that we can obtain.

The data we obtain comes from the various websites listed above. Looking at all the data, we believe that the discussions and comments on the topic of game development contained in Bilibili are the most suitable for sentiment analysis.

First of all, because station B gathers a large number of people to discuss this topic. This is also because station B, as an excellent video website in China, has attracted many people to browse. So in terms of quantity, the comments under the video of station b are sufficient. Secondly, the comments on station b are generally shorter. In contrast, there will be a longer analysis and expression on Zhihu and Weibo. Their content tends to be a bit more specific and complex, which makes natural language processing significantly more difficult. Moreover, many of these discussions are neutral and do not show obvious tendencies, which will also produce a certain error in the conclusions obtained.

So we finally decided to choose the comments of station B as the object of our final analysis.

We have selected videos on 10 topics, and the names are as follows. Among them, 4 are clearly in favor of the game and the development of the game industry, 4 are clearly opposed, and the remaining two are open topics.

Table 1-1 Video picklist

Video title	Video attitude
Game philosophy: When someone tells you that playing games is useless, how do you refute it?	In the tank
What do you think about the central media's article to support the development of the game industry? Looking at Genshin Impact's cultural output from the perspective of gamers, the domestic game industry is about to usher in spring?	In the tank
Psychoanalysis: What are we addicted to when we are addicted to games?	oppose
Objectively I admit that there are a lot of great games, but I really want to know: what is the growth that games have brought you?	oppose
The meaning, numbness and nihilism of video games	oppose
Did playing the game help you? What's the point of playing games? In-depth analysis! (Issue 0)	opening
Is a life without games a high-level, meaningful life?	opening
"It's a pity that you don't play the game and don't understand the meaning of this video"	In the tank
Are video games spiritual opium? 10 games that make people learn and make a name for video games	In the tank
Hua Wu Xueba is addicted to games, and Zhang Xuefeng talks about how he opens up	oppose

We got the comments on these 10 videos through the crawler. Each video was selected with the 500 comments with the highest number of likes, for a total of 5,000

comments as our final analysis dataset. Sorting by the number of likes is also a filter for the quality of the reviews, so that we can get higher quality reviews for analysis.

The 5,000 comments are the data we want to apply to the model generated by machine learning. That is, we hope to finally label these 5,000 pieces of data through machine learning, that is, to determine whether these comments are positive or negative. Then we can make a certain analysis of the conclusions. Next, we need to look at machine learning models.

1.3 LSTM&word2vec

1.3.1 RNN

Let's start with RNNs (Recurrent Neural Networks).

Why was RNN introduced? The reason is simple: when performing text processing, the previous text has an effect on the current part of speech of the text. In order to solve this probabilistic bias of traditional neural network models, we introduce RNNs.

The schematic diagram of the RNN is as follows:

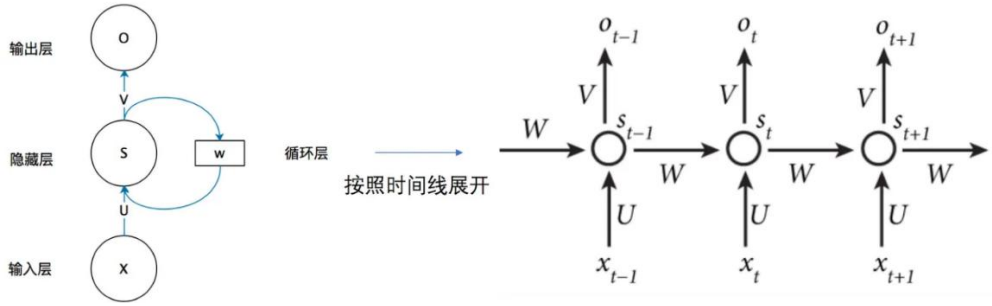

Figure 1- 1 Schematic diagram of an RNN

where x is the current input, o is the output, w is the weight matrix, and you and v are both weight matrices.

To put it simply, the state of s is not only generated by the change of the dot product of x and u, but also considered according to the processing of and w. This solves our problem: how to refer to the previous text state. When we get the current state s, we can get the output o by multiplying the dots by and v.s_{t-1}

The mathematical formula can be expressed in the following form:

$$O_t = g(V \cdot S_t)$$

$$S_t = f(U \cdot X_t + W \cdot S_{t-1})$$

where g and f are both handlers.

1.3.2 LSTM algorithm

Although the RNN algorithm is designed to have the function of short-term

memory, the problem is that as a comment text, we sometimes deal with some long sentences, and the memory ability of the RNN is insufficient. In addition, since RNNs need to be differentiated every time, there are problems with vanishing gradients and bursting gradients. So we're going to introduce a new algorithm, which is the LSTM algorithm.

LSTM stands for Long Short Term Memory, which is used to solve the long-term dependence problem that is common in general RNNs, and the use of LSTM can effectively transmit and express information in long-term series without causing useful information to be ignored for a long time. At the same time, the LSTM can also solve the problem of gradient vanishing/explosion in RNNs.

In the process of processing long texts, we will extract some keywords and understand and express the information based on them, which is what we need to store for a long time. Other texts can be stored as short-term memories. At the same time, we need to set up a dynamic mechanism to ensure that keywords are updated after they are no longer useful.

In a simple RNN algorithm, a separate tanh function is generally used:

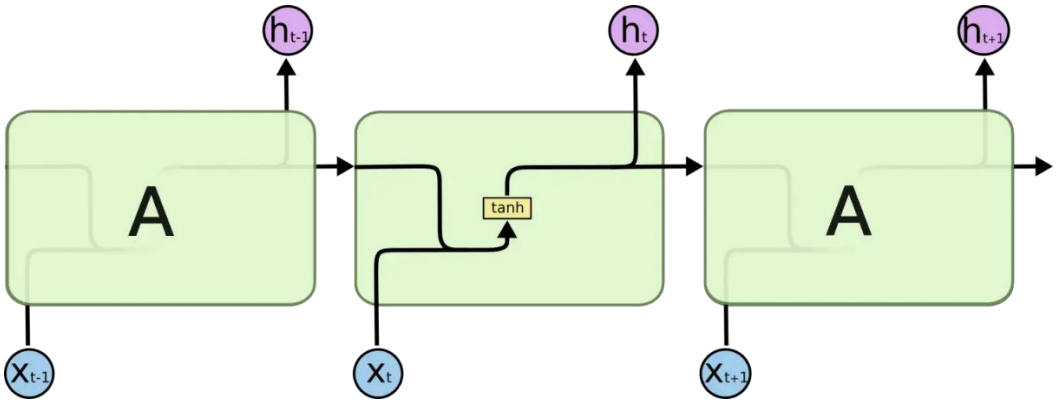

Figure 1-2 RNN schematic diagram 2

The network layer of the LSTM has four to interact with, which can be divided into three parts:

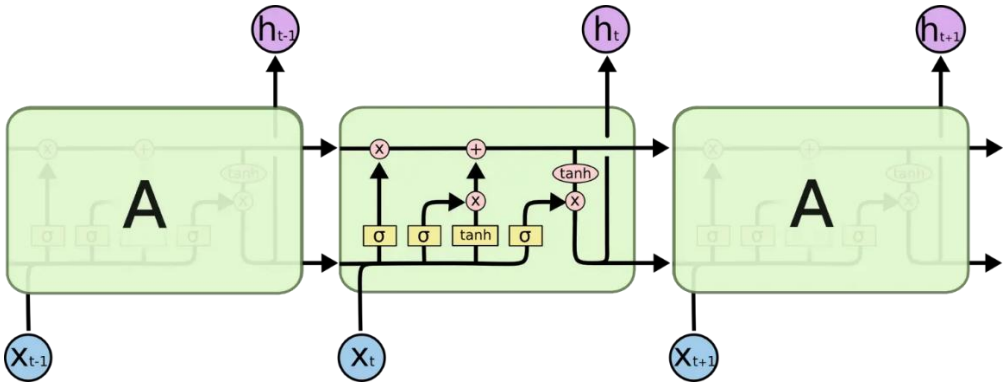

Figure 1-3 General diagram of the LSTM principle

The top of the network layer is called the cellular state, which only retains some linear interactions, so the information is easy to maintain on it, and this is used to

transmit information;

How to remove or add information to the cellular state is implemented through a gate, which is controlled by a sigmoid function (σ) and bitwise multiplication (). After gate processing, 1 means completely retained, 0 means discarded, and the middle one is in the intermediate state.×

The LSTM has three gates to control the state. One gate (left) determines whether to forget, one gate (middle) determines whether to update, and one gate (right) determines the output range.

The leftmost one is called the forgetting layer, and the formula is used to determine whether it is forgotten or not. $f_t = sigmoid(W_f \cdot [h_{t-1}, x_t] + b_f)$

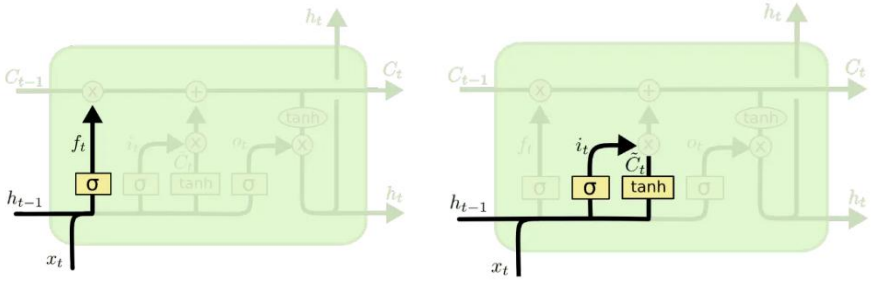

Figure 1-4 LSTM schematic part 1 Figure 2-44 LSTM schematic diagram 2

In the middle, you can add layers, as shown in Figure 2-5, and the following formula is used to determine whether to add layers:

$$i_t = sigmoid(W_i \cdot [h_{t-1}, x_t] + b_f)$$

The following formula is used to determine whether to update C:

$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

Combining the two, we get the following formula, as shown in Figure 2-6:

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t}$$

The left item is to determine whether to delete the original information, and the right item is to determine whether to update the new information, which is the update of the cell state.

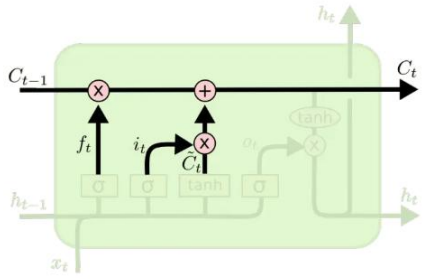

Figure 1-5 LSTM Schematic Diagram Part 3

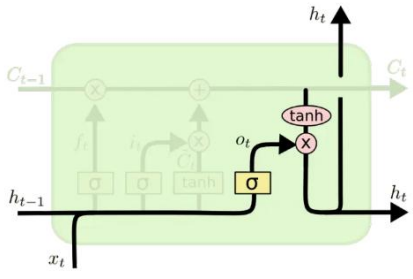

Figure 1-6 The principle of the LSTM

Finally, the last gate is used to process the output:

We use

$$o_t = sigmoid(W_o \cdot [h_{t-1}, x_t] + b_o)$$

Determine which parts need to be output (i.e., limit the output), and then use the tanh function to process the state and output the part of the instruction. In terms of formulas, it is:o_t

$$h_t = o_t * \tanh(C_t)$$

h_tThis is the output value we need, as shown in Figure 2-7, and the entire LSTM algorithm is over.

1.3.4 Word Embedding

After introducing the basic principles of the LSTM model, we found that there is still a big obstacle if we want to directly apply the neural network model to text sentiment analysis. Existing machine learning methods are often unable to process text data directly. This is because almost all machine semester models need to learn vector data. This requires finding a way to convert text data into numeric data.

The original idea was to make use of one-hot encoding, and the idea was that its basic idea was to represent each word as a binary vector. where the length of the vocabulary is the length of the vector. In a vector, if a word appears in the vocabulary, its word frequency is written in that position, and the rest of the positions are 0. Obviously, such a vector dimension is very high, and its disadvantages are obvious: first, the vector is too sparse and the learning efficiency is low. At the same time, he could not capture the relationship between words.

In order to better address these shortcomings, the concept of word embedding is derived.

Word embedding is natural language processing, which refers to embedding a high-dimensional space with the number of all words into a continuous vector space with a much lower dimension. Each word or phrase is mapped as a vector on a field of real numbers, which is also a distributed representation: each dimension of the vector has no practical meaning, while the whole represents a specific concept. The basic idea is to better capture the semantic relationships between words by representing them as dense vectors so that semantically similar words are closer together in the vector space.

Among them, word2vec technology is a kind of word embedding, which is more popular.

The word2vec technology was proposed by Google in 2013 to map words to vectors in a continuous space based on a neural network model. Word2Vec offers two training methods: Skip-gram and Continuous Bag of Words (CBOW). Skip-gram predicts the context words around it by giving a central word, whereas CBOW predicts the central word by giving the surrounding context words.

The core idea of the CBOW model is to predict target words based on context words, and its architecture typically includes an input layer, a projection layer, and an output layer. The number of units of the input layer is equal to the number of context words multiplied by the dimension of the word vector. The projection layer is a hidden

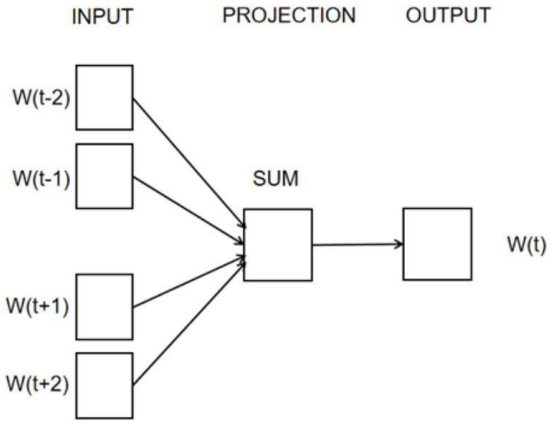

Figure 1-7 Schematic diagram of CBOW technology

layer between the input and output layers, and its dimension is often referred to as the dimension of the word vector. The number of units in the output layer is equal to the size of the vocabulary, each unit represents a vocabulary, and the probability distribution of the target word is output.

The skip-gram model, as opposed to CBOW, has the core idea of predicting context words based on target words. The architecture of the skip-gram model is similar to that of CBOW, but the dimension of the output layer becomes the size of the vocabulary. The number of cells in the input layer is equal to the size of the vocabulary,

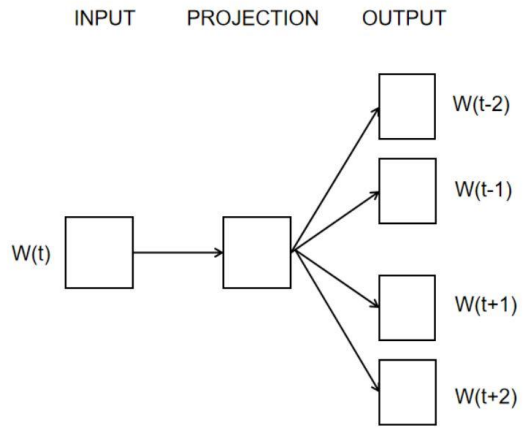

Figure 1-8 Schematic diagram of the Skip-Gram technique

each unit represents a vocabulary, and receives the one-hot encoding of the target word as input.

Word2vec has achieved some success in solving the problems of dimensional explosion, vector sparseness, and similarity between words, and we mainly choose this

method for word embedding.

Combining these models, we can convert text data into vector data through embedding technology, and then we can use LSTM models to learn and train them, and finally make predictions.

1.4 LSTM model training

1.4.1 Training Set

We want to get a dataset that includes the text data of the comment and the label that the comment has a positive or negative sentiment. With such a dataset, we can apply the techniques mentioned above for learning and training. Since we ended up analyzing 5,000 comments, we wanted the training set to have more data to achieve the effectiveness of the training. However, it should be admitted that the acquisition of the training set is more difficult. Since there was no existing dataset for labeled game discussions, we had to look for other approaches. We can manually mark and sort the collected labels. However, since hand-tagging is not very efficient, it is impossible to obtain too much data, so we prefer to use this part of the data as a test set rather than a training set. The final solution was to consolidate some of the collected data sets. First of all, we found the dataset of Weibo comments, weibo_senti_100k, a total of 100,000 comments, and we extracted 1,375 pieces of data related to the game according to the keyword extraction. Secondly, since we found that there were many reviews related to the specific content of the game in the dataset to be applied, such as the gameplay and game mechanics (such as criticizing krypton), we added 3625 reviews of the game on TapTap (TapTap is a game platform). But the number of these still seems to be a little insufficient, so we used ChatGPT to generate 5,000 comments on our own. We assign it to generate half positive reviews and half negative reviews, and give it a relevant topic, such as the dangers of negative addictive games, to generate review data on its behalf.

So we ended up generating 10,000 labeled comments for learning.

1.4.2 Stop word sets

using python39. The pre-process includes using the jieba library for Chinese word segmentation, establishing a list of stop words, and finally generating a dictionary. The deactivation glossary here is chosen hit_stopwords .txt (the specific source is given in the citations and references). Since all the comments are related to the game, we added "game" to the stop word.

At the same time, the file format of the training set is processed and prepared for learning.

1.4.3 hyperparameter configuration

word2vec and LSTM models. We use the embedding layer with dimension 128 and the lstm hidden layer, and use 3 LSTM layers. The drop rate is set to 0.8 to prevent

overfitting, and the learning rate is set to 0.0011 and the number of iterations is set to 100. This has been tested and found to work best. With these parameters configured, we train the model.

Table 1-2 Hyperparameter settings

Embed_size	Hidden_size	Layers_num	dropout	Learn_rate	Epochs_num
128	128	3	0.8	0.001	100
				1	

The following figure shows the change of the loss function during training. After 40 iterations, the value of the loss function tends to stabilize (the loss function of the later iteration process is not drawn).

Figure 1-9 Graph of loss function as a function of iteration

1.4.4 Test Set and Test Results

In the test set, we selected 1000 bilibili video reviews that we acquired and marked ourselves. Then we put the model to the test. Of the 1,000 pieces of data, there are 516 positive and 484 negative.

The results are presented using three values: accuracy, recall, and F1 score, as shown in the table below:

Table 1-3-Model test results

Accuracy	recall	F1_score
65.6%	70.9%	68.1%

At the same time, the confusion matrix for LSTM training is plotted:

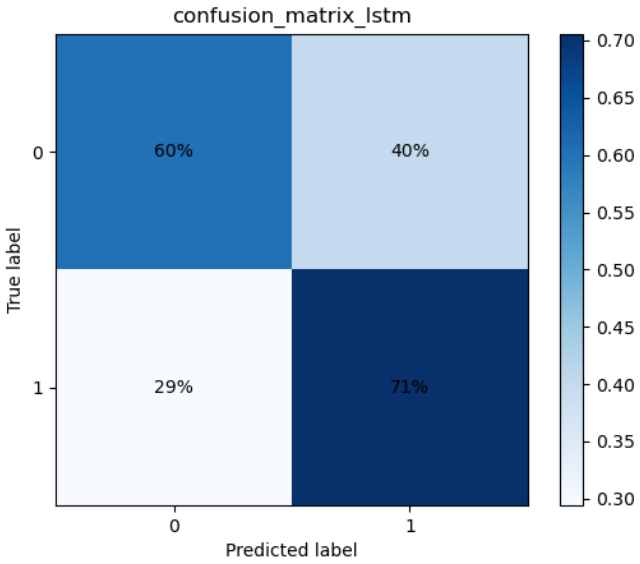

Figure 1-10 Confusion matrix

Overall, the model has some accuracy. Some features of the test set may not be covered due to some differences between the content of the training set and the test set. It may also be due to the accuracy of the hand-marked label itself, there is a certain error. In any case, the three indicators of this model are close to 70%, and it still achieves a good training effect.

1.5 LSTM application and result analysis

Next, we'll apply the model to crawl the comments of the bilibili ten videos. We treated each video comment separately and looked at the number of positive and negative emotions in each of them.

First, overall, a total of 3,065 of the 5,000 reviews were identified as positive by the model, accounting for 61.3%, while the remaining 39.7% were negative.

If we consider the accuracy of the model, combined with the accuracy and recall data of the model, that is, if there is only a 71% probability of each positive comment and only a 60% probability of a negative review, then the results can also be corrected as follows.

$$pos^* = pos * accuracy_{pos} + neg * (1 - accuracy_{neg})$$

The proportion of positive comments after correction, the proportion of positive and negative comments before correction is represented by POS and NEG, and the accuracy of prediction of positive samples and the accuracy of prediction of negative samples are represented, respectively. The revised result was 59.4 per cent. That is, after the amendment, it is believed that $pos^* accuracy_{pos} accuracy_{neg}$ 59.4% of the comments are still positive.

Secondly, we can also observe the distribution of the results of the comment data

for each video individually.

We selected the number of positive comments on each video as the proportion of the review summary obtained by this video as the characteristic quantity, and presented the results as follows:

Table-1-4 Percentage of positive comments in videos

Video title	Positive proportions
Game philosophy: When someone tells you that playing games is useless, how do you refute it?	69%
What do you think about the central media's article to support the development of the game industry? Looking at Genshin Impact's cultural output from the perspective of gamers, the domestic game industry is about to usher in spring?	61%
Psychoanalysis: What are we addicted to when we are addicted to games?	57%
Objectively I admit that there are a lot of great games, but I really want to know: what is the growth that games have brought you?	61%
The meaning, numbness and nihilism of video games	61%
Did playing the game help you? What's the point of playing games? In-depth analysis! (Issue 0).	65%
Is a life without games a high-level, meaningful life?	62%
"It's a pity that you don't play the game and don't understand the meaning of this video"	60%
Are video games spiritual opium? 10 games that make people learn and make a name for video games	62%
Hua Wu Xueba is addicted to games, and Zhang Xuefeng talks about how he opens up	55%

Combined with the emotional tendencies of these videos themselves, we draw the following histogram.

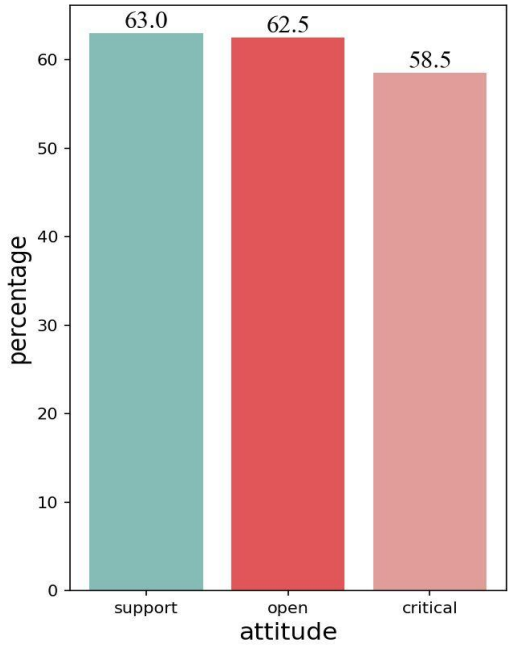

Figure 1-11 The relationship between the attitude of each video and the proportion of positive comments is indicated

This shows that when the UP master's video is in support of the development of the game or game industry, the emotional attitude of the comments will be more positive. This is in line with common sense, and it also shows the reliability of our model from the side.

To sum up, out of the 5,000 reviews we get, about sixty percent of the reviews are positive. This shows that most netizens support the development of games and the game industry. They don't have an aversion or positive attitude towards the game, which leads to our final conclusion.

This chapter first introduces the overall goals and ideas, and then starts with the purpose, introduces the review dataset that you want to obtain, and puts forward the sentiment analysis that you hope to obtain. Then, starting from the process, the relevant principles of the LSTM model are introduced, and the rationality of the goal is analyzed. Then, the principle and the target are connected, and the model is trained, and finally the LSTM model with an accuracy of 66% is obtained. Finally, it was applied to the review dataset, and the results obtained were that 59.4% of netizens had a positive attitude towards the development of games and the game industry after the correction.

2. Correlation analysis

2.1 Dataset Description

As already briefly described in the previous chapter, we collected tens of thousands of comments about video games from four platforms (Weibo, News, Zhihu, and Bilibili). The main purpose of this is to fully and comprehensively reflect the views of the "virtual nation" formed by all sectors of society on the Internet platform on the matter of "video games".

These reviews are long and short, rich in content and varied. The evaluation subjects are also highly diversified, from the perspective of industry, there are news media, online bloggers, educators, game industry workers, etc.; In terms of age, there are teenagers, young adults, middle-aged, and so on. And when selecting reviews, it is also based on a random sampling algorithm to ensure that there is no subjective error caused by human preference.

While the previous chapter focused on the general emotional attitudes we collected from the reviews, this chapter focuses on the focus and thinking perspectives of the reviews, and in a nutshell, the connections between video games and other things revealed in the reviews.

2.2 Principles of Statistics

Text is a discrete sequence of morphemes. This discrete sequence has a distinct statistical characteristic.

The first characteristic is structural. Sequences are limited by grammatical rules, and morphemes at different positions have different ideographic importances. This means that we can compress the segment based on word segmentation technology, remove redundancy, and make it concise in form and full of meaning. This is the text summarization technique. Based on the text summary, we were able to extract key information and ideas from a large number of comments, making it easier to understand and analyze the content of the review.

The second characteristic is aggregation. The sequence is limited by ideographic needs, and the morpheme ideographic themes that are spatially close are similar. This means that we can chunk segments based on keyword recognition technology, reducing dimensions in order to divide and conquer. That's the topic analysis technique. Through topic analysis, we can identify the emotional tendencies of the different topics involved in the review and classify them into different categories. Note that there are hierarchies on topics, video games are a topic, and there are different perspectives around video games in the commentary text, which are called sub-topics.

The third characteristic is labeling. The morphemes in the sequence naturally have a distinction between praise and disapproval, which is the embodiment of emotion and

the basis for labeling. Through sentiment analysis technology, we can evaluate and filter reviews based on their emotional tendencies to determine the final value orientation.

Therefore, statistics is the cornerstone of natural language processing technology, whether it is judging the importance of structure, screening complete sets of topics, or determining sentiment labels, all of which are based on the statistics of massive texts. The existing Python processing library has encapsulated the research results of a large number of scholars. However, due to the high complexity of real-world problems, the use of natural language technology to solve real-world problems still requires researchers to use them flexibly.

2.3. Data Preprocessing

We start by weeding out useless data.

Then, we screened about 500 positive comments and about 500 negative comments through text summarization and topic analysis techniques. This method aims to extract a representative sample from a large number of reviews for subsequent analysis and research.

2.4 TF-IDF

2.4.1 Principle

TF:

Word frequency refers to the frequency with which a word appears in document D, and is calculated as.

This formula indicates how often a word appears in a document.

IDF：

The inverse document frequency is the inverse number of the frequency of a word appearing in the entire document set.

The total number of documents in the document set D refers to the number of documents in the entire document set, and the number of documents containing the word t refers to the number of documents containing the word t.

TF-IDF value:

The TF-IDF value is the product of the word frequency and the inverse document frequency, which is used to comprehensively measure the importance of a word in a document, and is calculated as follows:

$$\mathrm{TF-IDF}(t, d, D) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t, D)$$

The calculation of the TF-IDF value will take into account the frequency of the word in the current document as well as the frequency in the overall document set, thus more accurately reflecting the importance of a word to the current document set. The principle is similar to that of information entropy, which reflects the information value of the words in the text set D. The greater the number of documents containing the word

t in the text set D, the lower its information value. If the number of words t in the text d is higher, the higher the information value. Obviously, this is a contradiction. Then the TF-IDF value comprehensively reflects the information value of the word by multiplication.

2.5 Thinking Perspective Statistics

2.5.1 Principle

After analyzing the high-value words of different platforms, we also analyzed the perspective of the reviewer's thinking, that is, the perspective from which the reviewer thinks about the game. This is a deeper layer in natural language processing – semantic analysis.

Semantic analysis aims to understand the meaning and intent behind a text, not just a superficial combination of words.

Topic model:

The topic model is a technique used to discover hidden topics in text, by identifying topics and topics in the text, we can infer the focus and concerns of reviewers, and infer the perspective from which they view the game.

Semantic Role Annotation:

Semantic role annotation is a technique of labeling semantic roles (such as agent, recipient, time, place, etc.) in a sentence, and by identifying semantic roles in text, we can infer the reviewer's attitudes and opinions towards different characters and elements in the game.

Semantic Similarity:

Semantic similarity is a measure of the semantic correlation between two texts, and by calculating the similarity between a commenter's comment and a text that is known to be semantically related (pre-filtered annotated text), it can be inferred whether the reviewer's thinking angle is similar to the known perspective. For example, a reviewer's review can be compared to an expert review or other user reviews to determine if their perspective is similar to that of others.

2.5.2 Visualization: Stacked histograms

Mental & Physical Fitness:

First, negative evaluations significantly discussed health problems, while positive evaluations discussed health problems less. This suggests that video games are more negative than positive for health. Especially for good health, positive evaluations basically do not start from this perspective, and there is a suspicion of avoiding the important and trivial.

The proportion of mental health topics on Bilibili is the highest, and Weibo is higher. Combined with the word cloud, it can be seen that on Bilibili, most of the groups

directly participate in the game, and they often discuss their personal feelings of playing the game. Weibo users are mostly parents, and they often discuss the health of their children's games.

There is a certain percentage of physical health discussions on Weibo, news and Zhihu, although less than Bilibili, but generally not low, reflecting people's concern about health issues.

Academic Performance:

First of all, both positive and negative comments discuss academic performance, and they are relatively rare, and similar, reflecting that the comments based on academic performance are relatively simple, and the positive and negative are evenly matched. Although this conclusion is more counter-real-life intuitive, there are indeed many people on the Internet who say that they can combine work and rest, or see games as a kind of entertainment, and do not think that video games have anything special to discuss for academic performance. However, the negative comments are more aggressive in terms of academic performance, and the language is very emotional.

As a knowledge-sharing platform, Zhihu has prominent discussions on academic performance, which may reflect users' concerns about learning methods and educational resources.

Work Career:

First of all, there were fewer positive and negative comments on academic performance, and slightly more positive comments.

The news media pays more attention to the work career, which is similar to the industry.

Family Relationships and Relationships:

First, there are more negative comments than positive comments about family relationships; However, when it comes to relationships, there are more positive comments than negative ones. This is a very interesting phenomenon. Combined with word clouds, it can be seen that video games are not conducive to parent-child communication, but they are conducive to gamers to socialize.

Weibo pays special attention to family relationships, which also proves that Weibo users are mostly parents and middle-aged people. On Zhihu, the discussion of interpersonal relationships is slightly higher. Family relationships and interpersonal relationships are discussed in a certain percentage on various platforms, but they are not the main topics.

Industry, industry and science and technology:

First of all, there are obviously more positive evaluations and discussions on industrial technology issues, while less negative evaluations are discussed. This shows that video games have more negative factors than positive factors for the industry, industry, and science and technology. Especially in science and technology, negative evaluations are basically not from this point of view, which is obviously justified given the GPU problem.

The prominence of the discussion of science and technology in the news reflects the media's focus on science and technology, which is actually a concern of the government, and we also focused on this issue in the first part.

3.References

[1] Zhang Shihong, Lai Degang, Huang Tingting. Research on Sentiment Analysis Algorithm Based on Recurrent Neural Network (RNN) in Natural Language Processing [J]. China Informatization, 2024, (03): 59-60+92.

[2] Hu Xianchen. Research and application of microblog text sentiment analysis method based on deep learning[D]. Nanchang University, 2023. DOI:10.27232/d.cnki.gnchu.2023.003439.

[3]colah. UnderstandingLSTMNetworks. (2015-08-27)[2024-05-14]. https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[4] Zhu Bin. Application of Improved Multi-Channel CNN-LSTM Model in Twitter Text Sentiment Analysis[D]. Henan University, 2019.

[5] Mikolov T ,0010 C K ,Corrado G , et al. Efficient Estimation of Word Representations in Vector Space [J]. CoRR, 2013, abs/1301.3781

4.appendix

4.1 Virtual Ethnography

Virtual ethnography is a relatively new concept that refers to a subject area that studies and describes ethnicities, races, or groups in the virtual world. With the development of virtual reality technology and the rise of virtual communities such as online games, people can create, shape, and experience a variety of different cultures, nationalities, and social groups in the virtual world.

The research scope of virtual ethnography covers the aspects of national identity, cultural inheritance, language customs, and social structure in the virtual world. Researchers can understand the formation and development of virtual nations by observing games, virtual social platforms, online forums, etc., and explore the connections and differences between virtual nations and real-world peoples.

The study of virtual ethnography can not only help people better understand the diversity and complexity of the virtual world, but also provide reference and enlightenment for cross-cultural communication, cultural communication and virtual community management. Research in this field often involves interdisciplinary collaborations, including knowledge and methods in the fields of sociology, anthropology, cultural studies, information science, and more.