IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 5, 2022, accepted February 20, 2022, date of publication March 8, 2022, date of current version March 18, 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3157715*

# A WeChat Official Account Reading Quantity Prediction Model Based on Text and Image Feature Extraction

**ZIJIAN BAI**[1], **SHUANGYI MA**[2], **AND GENG LI**[2]
[1]Tianjin Municipal Engineering Design and Research Institute Company Ltd., Tianjin 300392, China
[2]College of Management and Economics, Tianjin University, Tianjin 300072, China
Corresponding author: Zijian Bai (zijianbai2021@163.com)

**ABSTRACT** This paper describes a study that built a neural network prediction model based on feature extraction, focusing on text analysis and image analysis of WeChat official accounts reading quantity. Based on the embedding method of the deep learning model, we extracted the text features in the title and the image features in the cover picture, explored the relationship between these features and the reading quantity, and built a neural network model based on these features to predict the reading quantity. The results show that there is a phenomenon of sentiment fusion in the text, and a sentence vector model based on Doc2Vec and a neural network model both had a good performance. This paper proposes a tool that can predict the reading quantity in advance and help administrators adjust the titles and images according to the predicted results.

**INDEX TERMS** Feature extraction, neural network, WeChat official accounts, Doc2Vec, user engagement.

## I. INTRODUCTION

Social media platforms, such as Twitter and Facebook, provide opportunities for people to create, communicate, and share ideas. In China, WeChat is a social media platform with strong communication and influencing characteristics. Administrators apply for WeChat official accounts to publish different kinds of articles or news on the platform, and readers can obtain and share information. By November 2017, WeChat had gathered more than 10 million official accounts, including 3.5 million monthly active official accounts and 797 million monthly active users [1]. Many authors have formed their own brands through original articles and become entrepreneurs on WeChat.

Previous research on WeChat has focused on user behaviors and attitudes as well as the influence mechanism and communication power of WeChat as a social media platform. Specifically, it involves user satisfaction, user attitude, user intention [2]–[5], user engagement behavior [6], [7], and the influence mechanism and effect of WeChat as an information communication platform on service provided by users [8]–[11]. However, there has been little research on deeper mining and exploration of the text through natural language processing (NLP). Moreover, as far as we know, there has been little research on the analysis of the cover image of WeChat official accounts. Our study focuses on text analysis and image analysis of WeChat official accounts reading quantity, which contributes to the research in this specific field and addresses this research gap. Specifically, this study uses crawler technology to capture the WeChat official account data of the pet type. We extracted the text features in the title and the image features in the cover picture through the embedding method based on the deep learning model and explore the relationship between these features and the reading quantity. We also built a neural network model based on these features to predict the reading quantity of articles.

The study makes several contributions. Firstly, this paper is one of the first works to study the field of WeChat official accounts and effectively combine social media user engagement and multimedia elements. Secondly, a method is proposed to combine the text and image features of the title and cover image which achieves a breakthrough in text and image analysis in the field of WeChat official accounts. Moreover, the significance of our study is in the development of a tool that can predict the reading quantity in advance and help administrators adjust the titles and images according to the prediction. The results of this study can provide

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Tong.

administrators with feasible suggestions, help administrators better manage and improve WeChat official accounts, and expand the influence and communication power of accounts.

The remainder of this article is structured as follows: Section 2 provides a literature review that outlines some of the related previous studies. Section 3 introduces the research methods used in this study, followed by an analysis and discussion in Sections 4 and 5. Finally, we present the conclusion and future research directions in Section 6.

## II. LITERATURE REVIEW

This section briefly introduces current studies on social media engagement, multimedia elements, mainstream social media analysis methods, and WeChat official accounts.

### A. SOCIAL MEDIA ENGAGEMENT

Twitter, Facebook, and Instagram are typical social media platforms on which users can browse for information, express opinions, and follow or like accounts and posts. These user behaviors are forms of user engagement. Previous research has studied the three aspects of user engagement from the perspective of social media, namely, cognition, affect, and behavior [12]–[14]. Khan [15] summarized the forms of user engagement on social media platforms into two categories, participation and consumption, including like, comment, share, view, and so on. In another example of this, Devereux *et al.* [16] used the number of likes and comments as attributes of social media posts to study the engagement of small business consumers to better integrate marketing strategies. Moran *et al.* [17] measured the brand engagement of online consumers through Facebook participation indicators, namely clicks, likes, shares, and comments. Lim *et al.* [18] proposed a framework for studying social media engagement that focused on the influence of time and media type on user engagement. The research framework of Harrigan *et al.* [19] confirmed the multi-dimensionality of user engagement, and they found that three of them are of great significance to tourism social media engagement. Kim and Dennis [20] found that the credibility of an article affects user engagement in reads, likes, shares, and comments.

The importance of social media engagement has been continuously confirmed and discussed by previous research. However, the current research objects are more concentrated on foreign social media platforms, such as Facebook and Twitter, and little research has been conducted in detail regarding the social media engagement of WeChat official accounts [14], [21]–[23]. Therefore, this study focuses on social media user engagement of WeChat official accounts, taking reading as a form of user engagement, and we measure the impact of social media engagement through the index of the reading quantity.

### B. MULTIMEDIA ELEMENTS

Multimedia elements include text, video, sound, graphics, animation, and interactivity, all of which play an important role in social media [24]. A post on social media platforms consists of one or more multimedia elements, of which text elements are most used for social media analysis. For example, text elements in Twitter posts have been discussed to identify eyewitness information when a disaster occurred, and text elements in microblogs have been used to predict the stock market [25, [26]. Moreover, previous research has studied the influence of text elements on social media engagement and the close relationship between them [27]–[29]. Some studies have used the forms of user engagement in social media, such as likes and comments, as variables, to explore strategies or mechanisms. At the same time, the relationship between image and video elements in social media and user engagement has also been gradually studied by some scholars. Kordzadeh and Young [30] believed that the vividness of posts in multimedia elements like images and videos can help increase user engagement. Moran *et al.* [17] emphasized that the richness of media has a strong influence on all participative behaviors. They also found that visual images attract the most users to participate in behaviors.

### C. SOCIAL MEDIA ANALYTICAL METHODS

Social media analytics aims to collect, monitor, analyze, summarize and visualize social media data, and it is usually driven by specific requirements from a target application [31]. In recent years, social media analytics have been applied to politics [32]–[34], the economy [35], [36], culture [37], [38], the natural environment [39], health, and so on [40]. Commonly used social media analysis methods include topic modeling, semantic analysis, text classification, and sentiment analysis. Other machine learning methods, such as natural language processing, neural networks, and computer vision are gradually being applied to social media, making breakthroughs in the research methods. At the level of social media methods, the existing research reserves are very rich. Qiang *et al.* [41] proposed a geographical topic modeling method based on the Latent Dirichlet Allocation (LDA) model to find topics that start at a specific time. Munuswamy *et al.* [42] gave a new sentiment analysis rating prediction method and generated a new recommendation system. In the field of natural language processing, Kim *et al.* [43] used the Doc2Vec method to extract the feature vectors representing the technical meaning from the document text of an acquired company and estimated the company's technical similarity score with a start-up company. Sanz [44] used Doc2Vec to convert the report of each region into a numerical vector as a logistic regression feature to evaluate the sovereign credit risk of different regions. Furthermore, some research has combined social media and a neural network to achieve geographic location prediction [45].

Previous studies on WeChat have focused either on user behaviors and attitudes or on the influence of WeChat as a social media platform, and text analysis and image analysis of WeChat official accounts are very rare. Little research has been conducted on the deeper excavation and exploration of WeChat official accounts text through natural language processing. As far as we know, there is also very little research on the analysis of the cover image of WeChat official accounts. The title in the form of text and the cover image in the form of pictures are all the information of the article for users. The attractiveness of the text and pictures will directly affect the user's interest in clicking to read the article. How to adjust the title and cover image to achieve a higher reading quantity and improve the operating effect and dissemination of the WeChat official account has become a content worthy of research. With this as an entry point, this study started from the title and cover image of the article, and focused on the text analysis and image analysis of the WeChat official account reading quantity, which contributed to the research in this field and made up for this research gap.

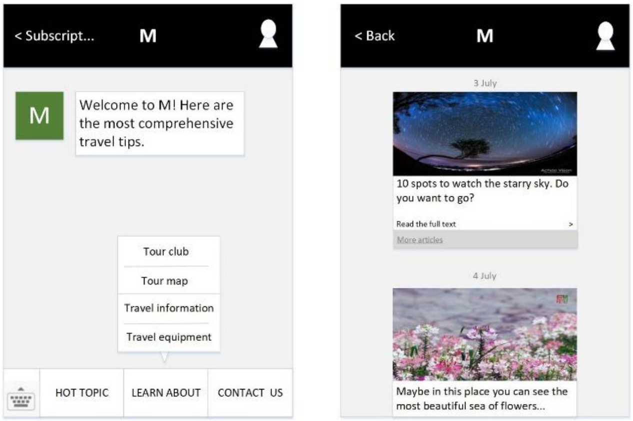
## III. WECHAT OFFICIAL ACCOUNTS

This section introduces the characteristics and importance of WeChat official accounts in detail.

There is a difference between WeChat and WeChat official accounts. WeChat is a cross-platform communication tool that was launched by Tencent in January of 2011. It supports single and multiple people to send voices, pictures, videos, and texts through a mobile phone network. However, WeChat is not only a chat tool but also a social media platform with strong communication and influence abilities. It provides functions like WeChat official accounts, WeChat Moments, and WeChat Pay. Users can send posts through WeChat Moments, add friends, and follow official accounts by searching for their number, scanning QR codes, and so on, and information can also be spread and shared through official accounts. By the second quarter of 2016, WeChat had covered more than 94% of China's smartphones, with 806 million monthly active users, reaching more than 200 countries and more than 20 languages.

As a function of WeChat, a WeChat official account is a service platform provided by WeChat for individuals and businesses. An individual or company can apply for an official account on the WeChat platform. After verifying identify information, the administrator of the account has the right to use the WeChat official account. Administrators can send, communicate, and share information through the official account, while users who follow the account can view information, participate in interactions, and give feedback through the official account. By November of 2017, WeChat had gathered more than 10 million official accounts, including 3.5 million monthly active official accounts [1]. The types of official accounts are diverse, including information inquiry, sharing, professional answers, interactive communication, and so on, involving education, tourism, security, news, governmental affairs, life, economy, and many other fields. The content of the article can be original or the integration of information.

A WeChat official account is divided into internal and external. In the internal interfaces, users can send messages to the official account, and the account will provide users with some functional modules in the menu bar at the bottom for users to select and view. The external interface displays the historical articles of the official account, all the articles that have been published will be displayed here, and the latest articles appear first. To more intuitively understand the difference between the two interfaces of WeChat official accounts, Figure 1 shows a simple figure of a tourist official account, which is shown in Figure 1. The picture on the left is the internal interface of the official account. Only users who follow the account can see this interface. The WeChat official account automatically sends a welcome message to the user and gives a brief introduction. At the bottom of the interface is a toolbar, and the far left is keyboard input. The user can input information when clicking, and the information will appear in this interface. The three modules on the right are the three functions provided by the official account for users, including ''Hot topic,'' ''Learn about,'' and ''Contact us.'' Among them, the second function includes four subfunctions, which further refines the needs of users. Each function is a hyperlink, which leads the user to enter a new link or obtain new information. The picture on the right is the external interface of the WeChat official account, and the historical articles are displayed in vertical order. The article format users see is composed of the title and cover image. Only when users choose to read the article can they click on the link to view the full text. The number of clicks on the article link is the page views of the article, which is also the reading quantity mentioned in this study. Furthermore, one official account can send multiple articles at a time, and the remaining articles will be folded.



Welcome to M! Here are the most comprehensive travel tips.

10 spots to watch the starry sky. Do you want to go?

Maybe in this place you can see the most beautiful sea of flowers...

FIGURE 1. The two interfaces of a WeChat official account.

## IV. RESEARCH METHODOLOGY

This section discusses the research framework used in this study.

This study focused on text analysis and image analysis on the reading quantity of WeChat official accounts. The research object selected in this study was the pet type, which is one of the different kinds of account types. Most of the ''protagonists'' in the pet-type articles are cats and dogs, and the WeChat official accounts selected in this study are of these two pets. The articles of the pet type have a wide range of content, including fun facts about pets, popularization of pet knowledge, etc. The data collected from the accounts was used to extract features through text analysis and image analysis. These data are described in the Data subsection. The study included the following steps:

- Acquire the title, cover picture, and the reading quantity of each article on the WeChat official accounts.
- Extract text features through natural language processing, such as sentiment analysis, text embedding, and other algorithms.
- Extract image features from the perspective of color and image recognition.
- Build a neural network based on these features and optimize the model to effectively predict the reading quantity.

### *A. DATA*

The pet-type official account was chosen as the research object because the official accounts under this type have many users, and the users view articles mainly based on curiosity and interest rather than an information platform that must be consulted, which was in line with our research purpose. More than 15 thousand data were collected from eight WeChat official accounts that were popular among the pet type, especially dogs and cats. The average reading quantity of these official accounts' articles ranged from 10 thousand to 70 thousand. We collected the title, cover picture, and link of each article as the original data, in which the title was represented as a text string, and the cover picture was represented as a link to the picture.

This study cleaned the original data through python, checked whether there were missing values under the three attributes of title, cover picture, and reading volume, and checked whether there were duplicate values under the attribute of title. After inspection, there was a duplicate item in the 15225 pieces of original data obtained, and there is no missing value. After data cleaning, there were 15224 data available in this study. In addition, since the need of subsequent text feature extraction, this study also segmented the text and removed stop words about the pre-processing of the data. After removing missing values and inappropriate data, the total amount of data collected was 15,224. These data were used to extract text and image features.

### *B. TEXT FEATURE EXTRACTION*

A total of 512 text features were extracted in this study, including the length of title, whether special symbols and numbers existed in the title, seven sentiment categories based on sentiment analysis, and 500-dimensional sentence vectors extracted based on text embedding. The detailed information of the text features is described in Figure 2. The special symbols here included "!," "?," and "…" The existence of these special symbols and the number in the title was a Boolean value, that is, if it existed, it was 1; otherwise, it was 0. Additionally, the seven sentiment categories were also Boolean values that represent the existence of certain sentiment words in the title. The 500-dimensional sentence vector based on text embedding is the expression of the text. Each dimension represents a word in the dictionary. Sentiment analysis and text embedding are discussed below.

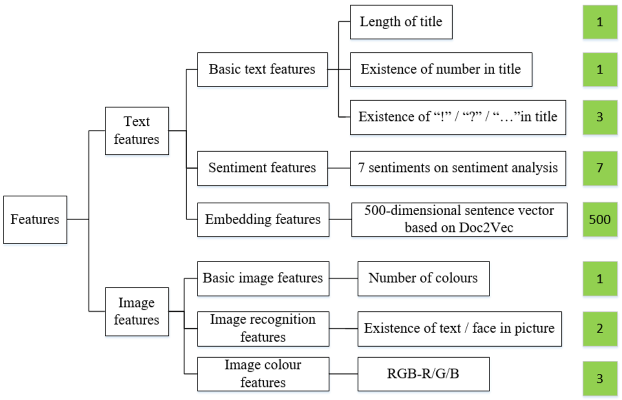

Features
Text features
Image features
Basic text features
Sentiment features
Embedding features
Length of title
Existence of number in title
Existence of "!" / "?" / "…"in title
7 sentiments on sentiment analysis
500-dimensional sentence vector based on Doc2Vec
1
1
3
7
500
Basic image features
Image recognition features
Image colour features
Number of colours
Existence of text / face in picture
RGB-R/G/B
1
2
3

**FIGURE 2. Detailed information of the text features and image features.**

#### 1) SENTIMENT ANALYSIS

Sentiment analysis is a common scenario in natural language processing including commodity evaluation, public opinion analysis, and sentiment classification, which plays an important role in guiding sentiment mining [46], [47]. Traditional sentiment analysis methods based on a sentiment dictionary are mostly used for public opinion analysis, and the sentiment tendencies are mainly positive and negative. However, the sentiment color of an article title is rich, and it is impossible to explore the sentiments contained in the sentence itself from only the positive and negative aspects.

The sentiment dictionary used in this study was the Chinese Sentiment Vocabulary Ontology Database of Dalian University of Technology, which is a Chinese Ontology resource organized and annotated by all the members of the Information Retrieval Research Office of Dalian University of Technology under the guidance of Professor Hongfei Lin. The dictionary describes Chinese vocabulary in terms of parts of speech, sentiment category, and sentiment intensity and polarity. Most importantly, the dictionary divides the sentiment color of vocabulary into seven categories: anger, disgust, fear, sadness, surprise, goodness and happiness. Compared with positive and negative sentiments, it greatly expands the range of sentiments. The above seven sentiment categories were taken as the seven variables of text features.

This study considered sentiment fusion, which refers to a collection of words with different sentiment colors in the title. Specifically, the dictionary lists the sentiment categories to which each word belongs, and we used the sentiment dictionary to find all the words with sentiment colors in the title of each article. There may have been several words in the title with different sentiments, or none. However, this does not mean the title had no sentiment colors but rather that the used words were not in the dictionary. The result of the sentiment categories is expressed in the form of a Boolean value, that is, if it existed, it was 1; otherwise, it was 0.

### 2) TEXT EMBEDDING

Unprocessed text cannot be used to quantitatively express the relationship and semantic information between texts. To achieve this, the text must be converted into a vector. The vector representation of words goes through the stage of transition for the one-hot encoding to distributed representation. The traditional one-hot encoding represents each word as an n-dimensional vector, where the length of the dimension is the size of the entire dictionary library, and each dimension represents a word in the dictionary. In the one-hot encoding representation of each word, the dimensions are all 0 except for one dimension which is 1. The dimension of 1 represents the word itself. However, any two words in this encoding method are isolated and cannot reflect semantic information. Moreover, it is easy to cause memory disasters because the dimensionality of the vocabulary is usually very large.

To solve this problem, the words are made into a distributed representation. Each word is mapped to a shorter word vector, and all the word vectors form a vector space. This process of embedding high-dimensional word vectors into a low-dimensional space is called word embedding. A word embedding is a type of text embedding that has gradually become an important part of natural language processing systems based on deep learning. They encode words and sentences in fixed-length vectors to greatly improve the processing performance of text data. The most used word embedding methods are Word2Vec and GloVe [48], [49]. The Word2Vec embedding method is a word vector computing tool launched by Google in 2013. It can vectorize all the words so that the relationship between words can be mined and measured quantitatively. Using the idea of machine learning, Word2Vec can simplify the text content into vector operations in the k-dimensional vector space through training, and the similarity in the vector space can be used to express the semantic similarity of text [50]. There are two models of Word2Vec, Continuous Bag of Words (CBOW) and Skip-gram. In the CBOW model, the word vectors of the context are cascaded or summed as features to predict the probability of the target word. The objective function is shown in Formula (1). Since the task of prediction is a multi-classification problem, the loss function uses SoftMax which is shown in Formula (2).

$$\frac{1}{T}\sum\nolimits_{t=k}^{T-k} logp\left(w_t \mid w_{t-k}, \ldots, w_{t+k}\right) \tag{1}$$

$$p\left(w_t \mid w_{t-k}, \ldots, w_{t+k}\right) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \tag{2}$$

Sentence and document embedding are the other forms of text embedding. Although Word2Vec can be used to represent a sentence vector, it ignores the influence of the sequence of words on the sentence of text information. Doc2Vec is an improvement of Word2Vec, which not only considers the semantics between words but also word order by adding a paragraph vector [48]. There are two models of Doc2Vec, Distributed Memory version of Paragraph Vector (PV-DM) and Distributed Bag of Words version of Paragraph Vector (PV-DBOW) [51].

In this study, we used the PV-DM model to extract sentence vectors. The model slides fixed-length words from a sentence at a time and takes one of the words as the predicted word, and the others as the input words. The word vector corresponding to the input words and the sentence-id vector serve as the input layer, and then the probability of the predicted word in the window is predicted. Window size refers to the maximum distance between the current word and the predicted word. For example, the current sentence could be ‘‘I played basketball last weekend’’ and the window size could be 2. In a certain slide, the prediction is ‘‘basketball,’’ and thus the input words are ‘‘I,’’ ‘‘played,’’ ‘‘last,’’ and ‘‘weekend.’’ The PV-DM model is shown in Figure 3.

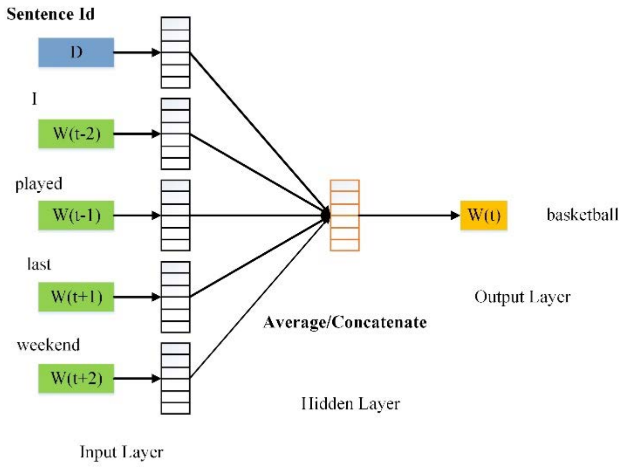


**FIGURE 3.** **The PV-DM model.**

It should be noted that, although the DM model is used to output the probability of the predicted word, the purpose of our study was to obtain the parameters in the model training process to obtain the sentence vector distribution expression of each text. In this study, the window size of the model was 3 and the sentence vector dimension was 500. To evaluate the performance of the model, we randomly selected a sample and set some sentences like the sample to measure the similarity between these sentences and the sample. The above process of randomly selecting samples was repeated many times. Then, we obtained the sentence vector distribution expressions of all the samples through the model. Each one was a 500-dimensional vector, which was the unique expression of sentence semantics, that is, the hidden layer vector in the model.

## C. IMAGE FEATURE EXTRACTION

A total of six image features were extracted in this study, including the number of colors, whether text and faces existed in the cover picture, and three red, green, and blue (RGB) val-

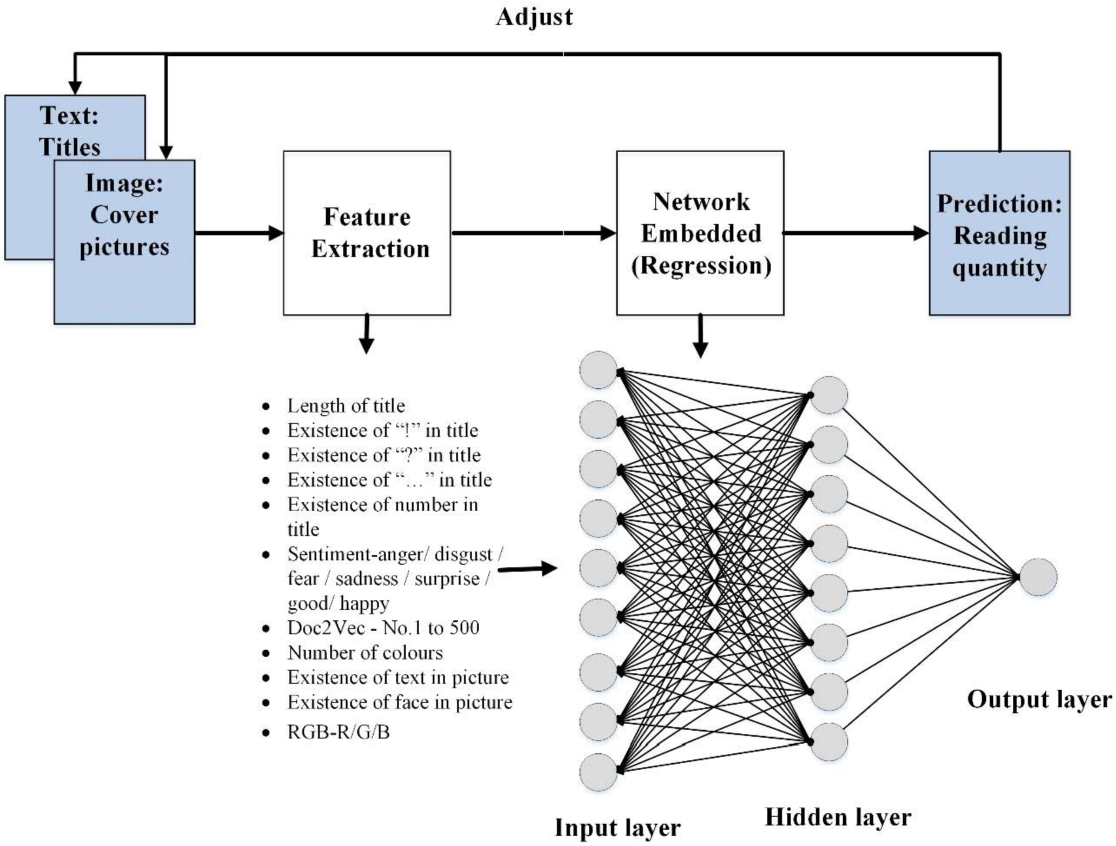

Adjust
Text: Titles
Image: Cover pictures
Feature Extraction
Network Embedded (Regression)
Prediction: Reading quantity
Length of title
Existence of "!" in title
Existence of "?" in title
Existence of "…" in title
Existence of number in title
Sentiment-anger/ disgust / fear / sadness / surprise / good/ happy
Doc2Vec - No.1 to 500
Number of colours
Existence of text in picture
Existence of face in picture
RGB-R/G/B
Input layer
Hidden layer
Output layer

**FIGURE 4.** The neural network based on regression prediction.

ues of the main color in the picture. The detailed information of the image features is described in Figure 2. The number of colors describes the richness of the colors in the picture. The higher the number, the richer the colors. The existence of the text in the picture was a Boolean value, that is, if it existed, it was 1; otherwise, it was 0. Additionally, the existence of the face in the picture was also a Boolean value. The existence of the text and face in picture was the result of image recognition. Due to the limited accuracy of the recognition method, our variables did not emphasize the specific situation of the recognition text and the number of faces but focused on whether it existed. The main color refers to the color with the largest proportion in the picture, which indirectly describes the dominant tone of the picture. This study used the RGB values of the main color.

### D. CONSTRUCTION OF THE NEURAL NETWORK

The previous steps obtained different text features and image features, which were used as the input variables of the neural network model, and the reading quantity was used as the output variable. It should be noted that, before building a neural network, all the features needed to be normalized because the order of magnitude difference between the features was very large. The variables to be normalized included the length of the title, number of colors, red (RGB-R), green (RGB-G), and blue (RGB-B). Moreover, the reading quantity was degraded by a factor of 100,000.

The neural network model used in this study was a regression prediction model, as shown in Figure 4. In this study, we used different features to predict the reading quantity of the article. Because the reading quantity is a continuous value, this problem is a regression problem, which is why we chose the regression prediction model. There were 518 input variables in the model, including 512 text features and the six image features mentioned above, and the output variable was the reading quantity of the article. The purpose of this study was to learn the impact of text and image features on the reading quantity, in other words, the impact of text and image elements on user engagement. This pattern is a way in which the title and cover picture are used to attract users to read the article. The result of the model can help administrators predict the reading quantity in advance and then make appropriate adjustments to the title and cover picture based on the predicted results to achieve a higher reading quantity.

We divided the data set into a training set, validation set, and test set. Specifically, the model was trained on the train set, the parameters were adjusted on the validation set, and then the model performance was evaluated through the test set. Although the eight WeChat official accounts were of the same type, each had its own unique attributes that affected

**TABLE 1.** Statistical results of sentiment analysis.

| Id | Sentiment - anger | Sentiment - disgust | Sentiment - fear | Sentiment - sadness | Sentiment - surprise | Sentiment - goodness | Sentiment - happiness |
|---|---|---|---|---|---|---|---|
| 30 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| 31 | 0 | 0 | **1** | 0 | 0 | 0 | 0 |
| 32 | 0 | **1** | 0 | **1** | 0 | 0 | 0 |
| 33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 34 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |

**TABLE 2.** Statistical result of sentiment words.

| Category | Representative words | Number of words |
|---|---|---|
| Goodness | Cute, like, friends, smart, health | 4106 |
| Disgust | Behind, mental-illness, breakdown, abandon, doubt | 3392 |
| Happiness | Happiness, joy, willingness, success, smile | 1101 |
| Sadness | Painful, innocent, regret, sorrow, lose | 936 |
| Fear | Careful, shock, fear, apology, harmful | 588 |
| Surprise | Unexpectedly, surprised, amazing, marvelous, miracle | 521 |
| Anger | Temper, punish, rear, complain, pique | 133 |

the accuracy of the neural network model results. To better realize the prediction of the reading quantity, we selected one of the WeChat official accounts as the data source of the neural network model. The official account had a total of 3,057 pieces of data. The result of the final data set division was 2,000 pieces of data in the training set, 500 pieces of data in the test set, and 557 pieces of data in the validation set. Among them, the ratio of the training set to the validation set was 4:1, and the ratio of the test set to the validation set was approximately 1:1. Finally, the mean square error (MSE), mean average error (MAE), and relative error were selected as the evaluation indicators to measure the performance of the regression neural network. The MAE and MSE are commonly used and typical evaluation indicators in regression problems. The relative error refers to the ratio of the absolute value of the error to the actual value, as shown in Formula (3). The smaller the relative error, the better the fitting effect. The parameters needed to be adjusted in the model include learning rate $\eta$, batch size, regularization parameter $\lambda$, number of layers, epoch, activation, and so on. The selection of parameters affects the training speed and fitting result of the neural network. For example, too high of a learning rate will lead to loss oscillation and a failure to converge, while too low of a learning rate will reduce the learning speed of the model. Moreover, to ensure that the learning results on the training data can be better applied to the test data, it is necessary to consider over-fitting and under-fitting and add some measures to suppress over-fitting like early stopping, regularization, and dropout strategy. In short, the selection and adjustment of parameters is a process of continuous testing. We needed to determine the appropriate parameters through the loss change during the training process and the results of the evaluation indicators.

$$Relative\ error = \frac{\sum_{i=0}^{n_{samples}} \left|\frac{y_{predict}-y}{y}\right|}{n_{samples}} \tag{3}$$

## V. ANALYSIS AND RESULTS

This section discusses the data analysis results of the above methods.

### A. SENTIMENT ANALYSIS RESULTS

The sentiment analysis results are shown in Table 1. It was found that the two sentiment words of disgust and sadness existed simultaneously in the 32nd sample, which was the expression of sentiment fusion in this study. The 30th, 31st, and 34th samples had only one sentiment word, while the 33rd sample did not detect any sentiment words. However, as previously discussed, this does not mean that the text had no sentiment color but rather that the words in the sample did not belong to the sentiment words in the dictionary. In addition to the sentiment distribution of each sample, we also counted the occurrence of the representative sentiment words in all the samples. Table 2 lists the number of each sentiment category and the top five representative sentiment words in each category. The results show that the words with sentiment colors of good and disgust existed the most times, which were 4106 and 3392, respectively. The sentiment words of anger existed the least, at only 133. Additionally, the top five sentiment words reflected the focus of the article under the seven sentiment categories. For example, in the good category, ''cute,'' ''smart,'' and ''health'' were keywords that often appeared in articles. ''Harmful'' and ''careful'' were found in the fear category, which indicates that the topic of these articles was to remind readers of what pets needed to pay attention to and what was bad for pets.

### B. THE TEXT EMBEDDING RESULTS

Table 3 describes the result of the text embedding feature through Doc2Vev between a sample and a set of similar sentences. The correlation between sentences decreases sequentially. Thus, the six similar sentences in Table 3 are increasingly less semantically relevant to the sample, and the prediction results of the model show that the similarity gradually decreased, which is consistent with this characteristic. Since the order of the results confirms to the actual situation, the performance of the sentence vector model was good. Table 4 shows the first, second, third, 499th, and 500th dimensional vectors.

**TABLE 3. The result between the sample and a set of similar sentences.**

| Sample | Similar sentence | Similarity |
|---|---|---|
| The man has roasted pork belly on the balcony for three months continuously ...To attract his own dog? | 1. The man has roasted pork belly on the balcony for three months ...To attract his own dog? | 0.8388 |
| | 2. The man has roasted pork belly for three months ...To attract his own dog? | 0.78851 |
| | 3. The man has baked sausages on the balcony for three months ...To attract his own dog? | 0.7572 |
| | 4. The man has roasted pork belly for three months...To? | 0.73542 |
| | 5. The man has baked sausages for three months...To? | 0.61082 |
| | 6. The man roasted the meat. | 0.50759 |

After feature extraction, this study process was divided into two steps. The first step studied the relationship between independent variables and dependent variables through correlation analysis and cluster analysis. Specifically, the relationship between text features, image features, and the reading quantity are discussed in the form of a heatmap through correlation analysis. Then, the reading quantity was divided into five categories through clustering analysis to find the characteristics of each category and the differences among different categories. The second step learned the features through a neural network to achieve the prediction of the reading quantity. Since there were 518 input variables, we cannot display them all here; however, some of the normalized data is described in Table 5.

### C. THE RESULTS OF CORRELATION ANALYSIS AND CLUSTERING ANALYSIS

The variable correlation results are shown in Figure 5. It is indicated that length of title, existence of ''…'' in title, and number in title are positively correlated with the reading quantity, while the existences of ''?'' and ''!'' in the title are not significantly correlated. Furthermore, text in picture, RGB-R, and RGB-G are negatively correlated with the reading quantity, while the number of colors, face in picture, and RGB-B are positively correlated. The number of cluster categories and corresponding categories are shown in Table 6. Among them, Category 3 had the highest reading quantity at 96,950 with 777 pieces of data, which should be used as a key case to study the characteristics of the articles. Category 1 had the higher reading quantity at 61,078 with 1388 pieces of data while category 4 has the medium reading quantity at 39,012 with 2272 pieces of data. Categories 0 and 2 have the lower and lowest reading quantity at 22,803 and 7701, respectively. There were 3093 pieces of data in Category 0 and 7764 pieces of data in Category 2 with the largest number, which means focus should be placed on the attributes of articles under the two categories to improve the reading quantity. The clustering results show that the proportion of the category with the highest reading quantity was only 5%, while the proportion of the categories with the lower and lowest reading quantity was as high as 71%. How to reduce the proportion of these two categories, improve the reading quantity, and give reasonable suggestions is one of the purposes of this study. To study the characteristics of each category and the differences among different categories, we calculated the mean values of different features in each category, respectively. For example, the mean value of the number of colors in Category 3 was 43,701.27, the mean value of sentiment-disgust in Category 1 was 0.22, while the mean value of the length of title in Category 2 was 21.07. Category 3 is excluded while studying the differences among categories since the proportion of the Category 3 was only 5%. Figure 6 shows a comparison of the variables under different categories. As the mean value of the length of title goes from high to low (shown in Figure 6 (a)) and the existence of text in pictures increases (shown in Figure 6 (b)), the reading quantity decreased (category 1>4>0>2). These results are consistent with the correlation results in Figure 5, that is, as the length of title increased and the existence of text in pictures decreased, the reading quantity showed an increasing trend.

### D. THE RESULTS OF THE NEURAL NETWORK MODEL

The results of the neural network model are shown in Figure 7, which describes the visualization of partial parameters selection, final parameters, and the evaluation indicators results. Figure 7 (a) shows the changes of train loss during 100 epochs under different learning rates when the other

**TABLE 4.** Text features of sentence vector.

| Id | Doc2Vec- No.1 | Doc2Vec- No.2 | Doc2Vec- No.3 | … | Doc2Vec- No.499 | Doc2Vec- No.500 |
|---|---|---|---|---|---|---|
| 1 | 0.10223 | -0.01777 | -0.07964 | … | 0.26774 | -0.23114 |
| 2 | 0.39489 | -0.01315 | 0.03171 | … | -0.15194 | 0.20247 |
| 3 | -0.08980 | 0.07287 | -0.25842 | … | 0.01226 | -0.16910 |
| 4 | -0.28936 | -0.19193 | -0.39874 | … | -0.07597 | -0.21663 |
| 5 | -0.15810 | -0.09384 | -0.13011 | … | 0.04239 | 0.07188 |
| 6 | -0.18148 | 0.04499 | 0.08971 | … | 0.28029 | -0.13192 |

**TABLE 5.** The normalized data.

| Id | Length of title | Existence of number | Number of colours | Text in picture | Face in picture | Sentiment-goodness | RGB-G | Doc2Vec-No.2 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.36923 | 0 | 0.06403 | 0 | 0 | 0 | 0.0039 | -0.01777 |
| 2 | 0.41538 | 0 | 0.12159 | 0 | 0 | 1 | 0.8627 | -0.01315 |
| 3 | 0.33846 | 1 | 0.18899 | 0 | 0 | 0 | 0.1529 | 0.07287 |
| 4 | 0.32308 | 0 | 0.1003 | 0 | 0 | 0 | 0.7568 | -0.19193 |
| 5 | 0.38462 | 0 | 0.15 | 0 | 0 | 0 | 0.8705 | -0.09384 |
| 6 | 0.33846 | 0 | 0.21963 | 0 | 0 | 0 | 0.5764 | 0.04499 |

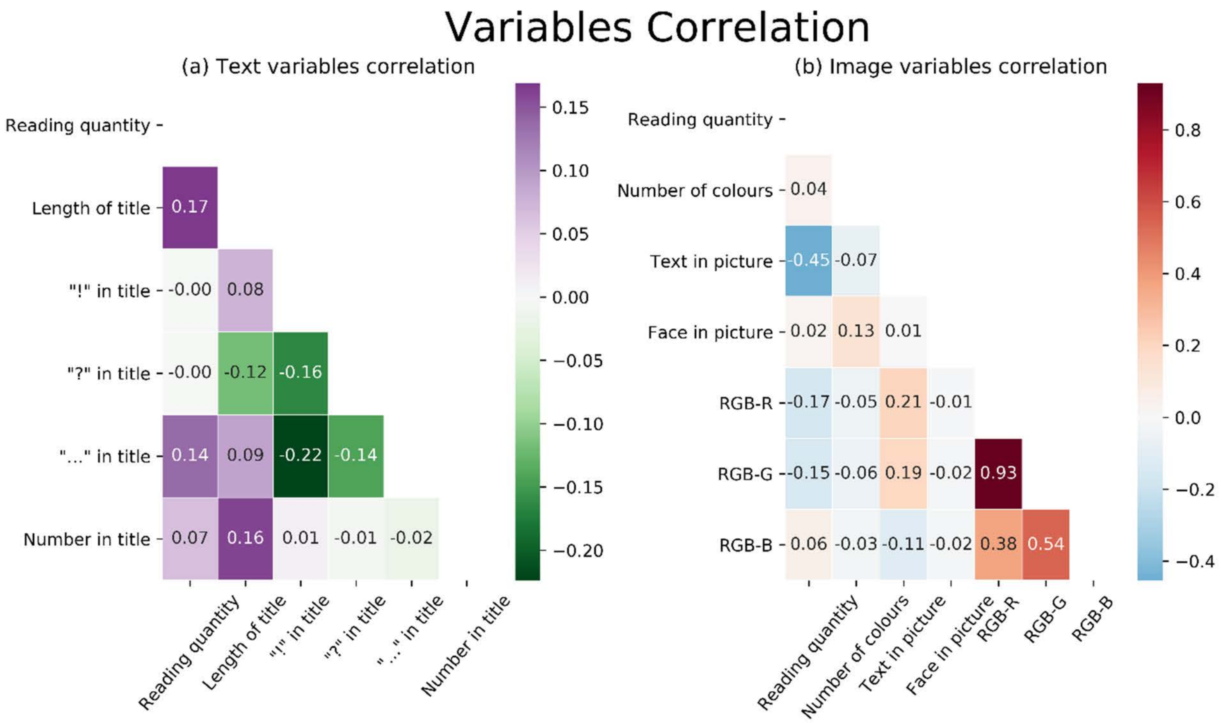


**FIGURE 5.** Variable correlation.

parameters were fixed. As a gradient explosion occurred when the learning rate was greater than 0.5, only some learning rates below 0.5 are drawn in the figure. As shown, as the learning rate increased, the training loss kept a trend of gradual decline and the became increasingly smaller. When the learning rate reached 0.25, the training loss dropped sharply

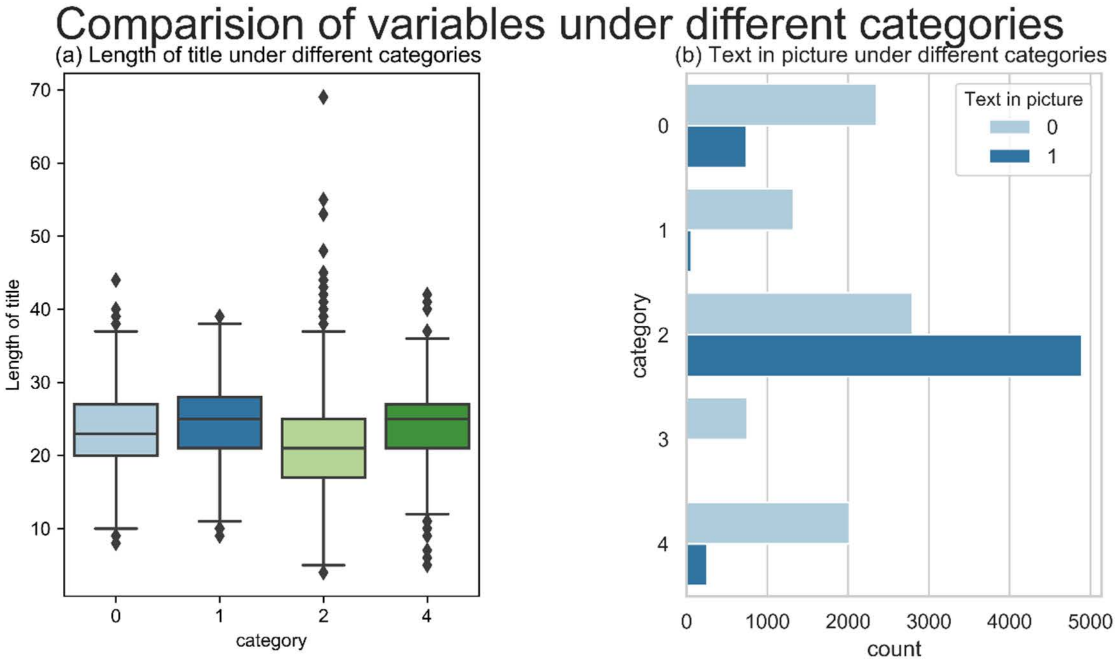


**FIGURE 6.** Comparison of variables under different categories.

**TABLE 6.** The result of clustering analysis.

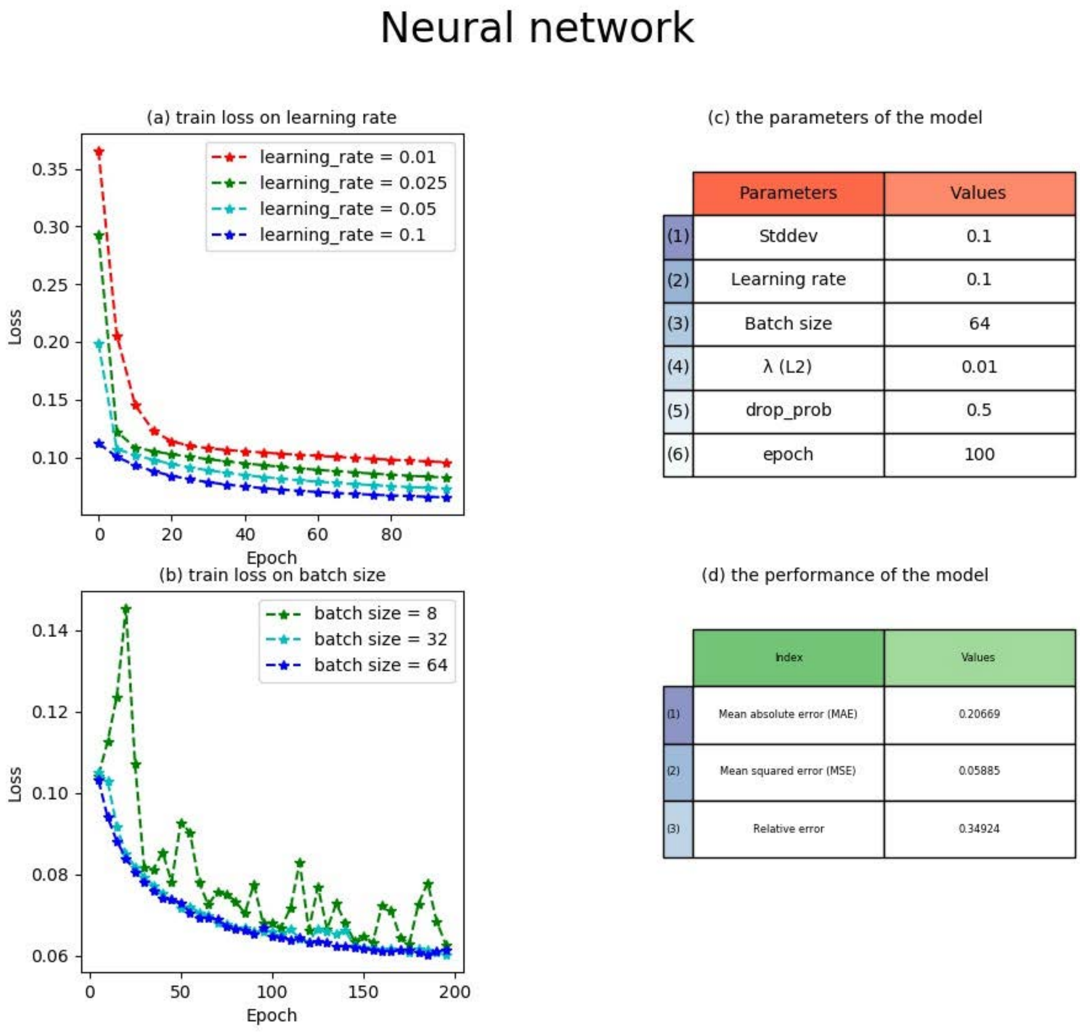
| Category | Reading quantity | Name of the category | Number of the category |
|---|---|---|---|
| 3 | 96950.37 | Highest | 777 |
| 1 | 61078.20 | Higher | 1388 |
| 4 | 39012.74 | Medium | 2272 |
| 0 | 22803.79 | Lower | 3093 |
| 2 | 7701.28 | Lowest | 7694 |

in the first few epochs, then slowly declined, and later reached saturation. The learning rate was finally fixed to 0.1 according to the threshold of the learning rate and the actual situation. Figure 7 (b) shows the changes of training loss of different batch sizes during 200 epochs. The selection of batch size also affected the speed and performance of the model. Too small a batch size may cause loss oscillation and a failure to converge, which is shown as the green curve in the figure; too large of batch size will increase the time cost. The changes of the other two curves were compared, in which the cyan curve represents the batch size of 32 and the blue represents the batch size of 64. Their training loss eventually drops to the same size and the two curves had a slight vibration during the whole epoch. Finally, the batch size was fixed to 64 in this study. Figure 7 (c) describes all the parameters that were finally determined. The standard deviation of the normal distribution of the weight initialization was 0.1, the learning rate $\eta$ was 0.1, and the batch size was 64 as mentioned above. The regularization parameter $\lambda$ was 0.0001, drop probability was 0.5, and the epoch was 200. Besides, the optimizer used in the model was a gradient descent optimizer. Figure 7 (d) describes the results of the evaluation indicators under the above parameters. The results show that the MAE was 0.207, the MSE was 0.059, and the relative error was 34.9%, which indicates that the model had a good performance.

## VI. DISCUSSION

The results reveal that there was a phenomenon of sentiment fusion in the titles of some samples, in which the words with sentiment colors of good and disgust existed the most times, while the sentiment words of anger existed the least. To further explore the expression forms of sentiment colors, we counted the top five representative sentiment words that occurred the most frequently in each sentiment category. The results show that “cute,” “smart,” “healthy,” and other words associated with good characters and the health of pets were more popular with the administrators, who published a very high number of articles in this category. Surprisingly, disgust, as a negative sentiment category, ranked second only to good in terms of the number of sentiment words. The findings reveal that “breakdown,” “abandon,” and “doubt” were the most common words in this category, which indicate that administrators pay more attention to the stories with this kind of sentiment and share them with readers.

This study also explored the sentence vector that can be used as a unique textual semantic representation. In this study, each text was extracted as a 500-dimensional vector, and each dimension represented a lexical word in the dictionary. The result of the text embedding based on Doc2Vec shows that



|  | Parameters | Values |
|---|---|---|
| (1) | Stddev | 0.1 |
| (2) | Learning rate | 0.1 |
| (3) | Batch size | 64 |
| (4) | λ (L2) | 0.01 |
| (5) | drop_prob | 0.5 |
| (6) | epoch | 100 |

|  | Index | Values |
|---|---|---|
| (1) | Mean absolute error (MAE) | 0.20669 |
| (2) | Mean squared error (MSE) | 0.05885 |
| (3) | Relative error | 0.34924 |

**FIGURE 7. Results of the neural network.**

our sentence vector model had a good performance in relative similarity and could predict the semantic information of new texts. This is probably because the research object of our official accounts was fixed to the pet type, and a single type is more concentrated in the use of words as well as being easier to build a dictionary of the same type.

The results of correlation and cluster analysis indicate that the increase of length of title as well as the existence of "…" in the title were more likely to promote user engagement and thereby increase the reading quantity of pet articles, but the existence of text in a picture may inhibit user engagement and reduce users' interests in reading, thereby decrease the reading quantity of articles. However, other features did not show a significant correlation, which may be due to the difference in styles of different WeChat official accounts. Moreover, the study found that the proportion of articles with the highest reading quantity was only 5%, while the proportion with the lower and lowest reading quantity was as high as 71%, which reveal that there was still a large difference in the reading quantity among samples. We hope to help administrators reduce the proportion of these two categories and increase the reading quantity.

Therefore, we proposed a neural network prediction model based on text and image features, which makes up for the lack of basic data analysis and can effectively help administrators to predict the reading quantity in advance through machine learning and adjust the title and cover picture according to the predicted results to achieve higher reading quantity. This study adjusted and optimized different parameters and added some approaches to suppress over-fitting. The more appropriate parameters are eventually determined through the change of training loss during the training process and the results of evaluation indicators. However, the model results of the regression neural network must be viewed cautiously given the high level of relative error (34.9%).

### A. LIMITATIONS AND FUTURE RESEARCH

This study had certain limitations that provided opportunities for future research. First, we only applied the proposed model to the pet type, whereas each type of WeChat official account may have its own style and development in the process of article sharing and publishing. Future research can expand the investigation of other types of WeChat official accounts and further explore the characteristics of articles in different fields. Second, there are many forms of user engagement in social media, of which only one was discussed in this study, namely, reading quantity. However, different forms of user engagement have different communication power and influence on social media platforms. Future research can try to explore the influence of multimedia elements on other user engagement behaviors, such as share and comment. Furthermore, we proposed a neural network model based on text and image feature extraction, which is a preliminary attempt to provide a prediction approach for administrators. However, both image feature extraction and model algorithms have room for improvement. Future research can delve deeper into more advanced methods and further extract image features to enrich the independent variables and improve the data diversity and accuracy of the model itself. In the selection of prediction models, algorithms like a convolutional neural network can be combined to achieve a performance comparison among algorithms.

## VII. CONCLUSION

This study effectively combined social media user engagement and multimedia elements, explicated the factors that may affect the reading quantity of WeChat official accounts articles through text analysis and image analysis, and proposed a neural network prediction model based on text features and image features. By collecting the text and image information from eight pet WeChat official accounts, we extracted 518 features through sentiment analysis and sentence vector and image recognition, and finally embedded them into the machine learning model to achieve this goal.

The results show that both the sentence vector model based on Doc2Vec and the neural network model had a good performance. This article proposed a comprehensive tool that can help administrators predict the reading quantity in advance and make appropriate adjustments to the title and cover picture according to the predicted results to guide administrators to better manage and improve the WeChat official accounts, attract the attention of readers, and expand the communication power and influence of the accounts. Moreover, the results of this study provide a strong basis for future research.

## REFERENCES

[1] Chatterbox. (2017). *The 2017 WeChat Data Report | WeChat Blog: Chatterbox*. WeChat. [Online]. Available: http://blog.wechat.com/2017/11/09/the-2017-wechat-data-report/

[2] C. H. Lien and Y. Cao, "Examining WeChat users' motivations, trust, attitudes, and positive word-of-mouth: Evidence from China," *Comput. Hum. Behav.*, vol. 41, pp. 104–111, Dec. 2014, doi: 10.1016/j.chb.2014.08.013.

[3] C. Gan and W. Wang, "Uses and gratifications of social media: A comparison of microblog and WeChat," *J. Syst. Inf. Technol.*, vol. 17, no. 4, pp. 351–363, Nov. 2015, doi: 10.1108/JSIT-06-2015-0052.

[4] C.-H. Lien, Y. Cao, and X. Zhou, "Service quality, satisfaction, stickiness, and usage intentions: An exploratory evaluation in the context of WeChat services," *Comput. Hum. Behav.*, vol. 68, pp. 403–410, Mar. 2017, doi: 10.1016/j.chb.2016.11.061.

[5] C.-B. Zhang, Y.-N. Li, B. Wu, and D.-J. Li, "How WeChat can retain users: Roles of network externalities, social interaction ties, and perceived values in building continuance intention," *Comput. Hum. Behav.*, vol. 69, pp. 284–293, Apr. 2017, doi: 10.1016/j.chb.2016.11.069.

[6] X. Cao, M. Gong, L. Yu, and B. Dai, "Exploring the mechanism of social media addiction: An empirical study from WeChat users," *Internet Res.*, vol. 30, no. 4, pp. 1305–1328, May 2020, doi: 10.1108/INTR-08-2019-0347.

[7] C. Gan, "Understanding WeChat users' liking behavior: An empirical study in China," *Comput. Hum. Behav.*, vol. 68, pp. 30–39, Mar. 2017, doi: 10.1016/j.chb.2016.11.002.

[8] X. Zhang, D. Wen, J. Liang, and J. Lei, "How the public uses social media wechat to obtain health information in China: A survey study," *BMC Med. Informat. Decis. Making*, vol. 17, no. S2, Jul. 2017, doi: 10.1186/s12911-017-0470-0.

[9] W. Li, L. Q. Han, Y. J. Guo, and J. Sun, "Using WeChat official accounts to improve malaria health literacy among Chinese expatriates in niger: An intervention study," *Malaria J.*, vol. 15, no. 1, Dec. 2016, doi: 10.1186/s12936-016-1621-y.

[10] H. Pang, "Unraveling the influence of passive and active WeChat interactions on upward social comparison and negative psychological consequences among university students," *Telematics Informat.*, vol. 57, Mar. 2021, Art. no. 101510, doi: 10.1016/j.tele.2020.101510.

[11] T. Jiang, Y. Wang, T. Lin, and L. Shangguan, "Evaluating Chinese government WeChat official accounts in public service delivery: A user-centered approach," *Government Inf. Quart.*, vol. 38, no. 1, Jan. 2021, Art. no. 101548, doi: 10.1016/j.giq.2020.101548.

[12] L. Dessart, C. Veloutsou, and A. Morgan-Thomas, "Consumer engagement in online brand communities: A social media perspective," *J. Product Brand Manage.*, vol. 24, no. 1, pp. 28–42, Mar. 2015, doi: 10.1108/JPBM-06-2014-0635.

[13] L. Hollebeek, "Exploring customer brand engagement: Definition and themes," *J. Strategic Marketing*, vol. 19, no. 7, pp. 555–573, Dec. 2011, doi: 10.1080/0965254X.2011.599493.

[14] D. Lee, K. Hosanagar, and H. S. Nair, "Advertising content and consumer engagement on social media: Evidence from Facebook," *Manage. Sci.*, vol. 64, no. 11, pp. 5105–5131, Nov. 2018, doi: 10.1287/mnsc.2017.2902.

[15] M. L. Khan, "Social media engagement: What motivates user participation and consumption on YouTube?" *Comput. Hum. Behav.*, vol. 66, pp. 236–247, Jan. 2017, doi: 10.1016/j.chb.2016.09.024.

[16] E. Devereux, L. Grimmer, and M. Grimmer, "Consumer engagement on social media: Evidence from small retailers," *J. Consum. Behav.*, vol. 19, no. 2, pp. 151–159, Mar. 2020, doi: 10.1002/cb.1800.

[17] G. Moran, L. Muzellec, and D. Johnson, "Message content features and social media engagement: Evidence from the media industry," *J. Product Brand Manage.*, vol. 29, no. 5, pp. 533–545, Oct. 2019, doi: 10.1108/JPBM-09-2018-2014.

[18] M. Kim, "Do media type and time of day matter in social media engagement? The case of the music industry," *Int. Telecommun. Policy Rev.*, vol. 24, no. 1, pp. 105–124, 2017.

[19] P. Harrigan, U. Evers, M. Miles, and T. Daly, "Customer engagement with tourism social media brands," *Tourism Manage.*, vol. 59, pp. 597–609, Apr. 2017, doi: 10.1016/j.tourman.2016.09.015.

[20] A. Kim and A. R. Dennis, "Says who? The effects of presentation format and source rating on fake news in social media," *MIS Quart.*, vol. 43, no. 3, pp. 1025–1039, Jan. 2019, doi: 10.25300/MISQ/2019/15188.

[21] R. Leung, M. Schuckert, and E. Yeung, "Attracting user social media engagement: A study of three budget airlines Facebook pages," in *Information and Communication Technologies in Tourism 2013*. Berlin, Germany: Springer, 2013, pp. 195–206.

[22] J. P. D. Guidry, Y. Jin, C. A. Orr, M. Messner, and S. Meganck, "Ebola on Instagram and Twitter: How health organizations address the health crisis in their social media engagement," *Public Relations Rev.*, vol. 43, no. 3, pp. 477–486, Sep. 2017, doi: 10.1016/j.pubrev.2017.04.009.

[23] S. L. Lo, D. Cornforth, and R. Chiong, "Use of a high-value social audience index for target audience identification on Twitter," in *Proc. Australas. Conf. Artif. Life Comput. Intell.* Cham, Switzerland: Springer, 2015, pp. 323–336.

[24] L. Y. B. Khedif, A. Engkamat, and S. Jack, "The evaluation of users' satisfaction towards the multimedia elements in a courseware," *Proc.-Social Behav. Sci.*, vol. 123, pp. 249–255, Mar. 2014, doi: 10.1016/j.sbspro.2014.01.1421.

[25] K. Zahra, M. Imran, and F. O. Ostermann, "Automatic identification of eyewitness messages on Twitter during disasters," *Inf. Process. Manage.*, vol. 57, no. 1, Jan. 2020, Art. no. 102107, doi: 10.1016/j.ipm.2019.102107.

[26] A. Sun, M. Lachanski, and F. J. Fabozzi, "Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction," *Int. Rev. Financial Anal.*, vol. 48, pp. 272–281, Dec. 2016, doi: 10.1016/j.irfa.2016.10.009.

[27] C. Ding, H. K. Cheng, Y. Duan, and Y. Jin, "The power of the 'like' button: The impact of social media on box office," *Decis. Support Syst.*, vol. 94, pp. 77–84, Feb. 2017, doi: 10.1016/j.dss.2016.11.002.

[28] K. Swani and L. I. Labrecque, "Like, comment, or share? Self-presentation vs. brand relationships as drivers of social media engagement choices," *Marketing Lett.*, vol. 31, nos. 2–3, pp. 279–298, Sep. 2020, doi: 10.1007/s11002-020-09518-8.

[29] Y. Chang, Y. Li, J. Yan, and V. Kumar, "Getting more likes: The impact of narrative person and brand image on customer–brand interactions," *J. Acad. Marketing Sci.*, vol. 47, no. 6, pp. 1027–1045, 2019, doi: 10.1007/s11747-019-00632-2.

[30] N. Kordzadeh and D. K. Young, "How social media analytics can inform content strategies," *J. Comput. Inf. Syst.*, vol. 62, no. 1, pp. 128–140, Jan. 2022, doi: 10.1080/08874417.2020.1736691.

[31] D. Zeng, H. Chen, R. Lusch, and S.-H. Li, "Social media analytics and intelligence," *IEEE Intell. Syst.*, vol. 25, no. 6, pp. 13–16, Nov. 2010, doi: 10.1109/MIS.2010.151.

[32] S. Stieglitz and L. Dang-Xuan, "Social media and political communication: A social media analytics framework," *Social Netw. Anal. Mining*, vol. 3, no. 4, pp. 1277–1291, Dec. 2013, doi: 10.1007/s13278-012-0079-3.

[33] Z. Tufekci and C. Wilson, "Social media and the decision to participate in political protest: Observations from tahrir square," *J. Commun.*, vol. 62, no. 2, pp. 363–379, Apr. 2012, doi: 10.1111/j.1460-2466.2012.01629.x.

[34] B. Kalsnes, A. H. Krumsvik, and T. Storsul, "Social media as a political backchannel: Twitter use during televised election debates in Norway," *Aslib J. Inf. Manage.*, vol. 66, no. 3, pp. 313–328, May 2014, doi: 10.1108/AJIM-09-2013-0093.

[35] P. Y. C. Kusuma, S. Sumpeno, and A. D. Wibawa, "Social media analysis of BPS data availability in economics using decision tree method," in *Proc. 1st Int. Conf. Inf. Technol., Inf. Syst. Electr. Eng. (ICITISEE)*, Aug. 2016, pp. 148–153, doi: 10.1109/ICITISEE.2016.7803064.

[36] E. A. Stoica, A. G. Pitic, and L. Mihăescu, "A novel model for E-business and E-government processes on social media," *Proc. Econ. Finance*, vol. 6, no. 13, pp. 760–769, 2013, doi: 10.1016/s2212-5671(13)00200-1.

[37] I. Pentina, L. Zhang, and O. Basmanova, "Antecedents and consequences of trust in a social media brand: A cross-cultural study of Twitter," *Comput. Hum. Behav.*, vol. 29, no. 4, pp. 1546–1555, Jul. 2013, doi: 10.1016/j.chb.2013.01.045.

[38] D. Ray, "Overcoming cross-cultural barriers to knowledge management using social media," *J. Enterprise Inf. Manage.*, vol. 27, no. 1, pp. 45–55, Feb. 2014, doi: 10.1108/JEIM-09-2012-0053.

[39] J. B. Houston, J. Hawthorne, M. F. Perreault, E. H. Park, M. Goldstein Hode, M. R. Halliwell, S. E. Turner McGowen, R. Davis, S. Vaid, J. A. McElderry, and S. A. Griffith, "Social media and disasters: A functional framework for social media use in disaster planning, response, and research," *Disasters*, vol. 39, no. 1, pp. 1–22, Jan. 2015, doi: 10.1111/disa.12092.

[40] A. Pentescu, I. Cetină, and G. Orzan, "Social media's impact on healthcare services," *Proc. Econ. Finance*, vol. 27, no. 15, pp. 646–651, 2015, doi: 10.1016/s2212-5671(15)01044-8.

[41] S. Qiang, Y. Wang, and Y. Jin, "A local-global LDA model for discovering geographical topics from social media," in *Proc. Asia–Pacific Web (APWeb) Web-Age Inf. Manage. (WAIM) Joint Conf. Web Big Data*. Cham, Switzerland: Springer, 2017, pp. 27–40.

[42] S. Munuswamy, M. S. Saranya, S. Ganapathy, S. Muthurajkumar, and A. Kannan, "Sentiment analysis techniques for social media-based recommendation systems," *Nat. Acad. Sci. Lett.*, vol. 44, no. 3, pp. 281–287, Jun. 2021, doi: 10.1007/s40009-020-01007-w.

[43] H. J. Kim, T. S. Kim, and S. Y. Sohn, "Recommendation of startups as technology cooperation candidates from the perspectives of similarity and potential: A deep learning approach," *Decis. Support Syst.*, vol. 130, Mar. 2020, Art. no. 113229, doi: 10.1016/j.dss.2019.113229.

[44] I. P. Sanz, "Using the European commission country recommendations to predict sovereign ratings: A topic modeling approach," *Expert Syst. Appl., X*, vol. 5, Apr. 2020, Art. no. 100026, doi: 10.1016/j.eswax.2020.100026.

[45] I. Lourentzou, A. Morales, and C. Zhai, "Text-based geolocation prediction of social media users with neural networks," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 696–705, doi: 10.1109/BigData.2017.8257985.

[46] Y. Lyu, J. C.-C. Chow, and J.-J. Hwang, "Exploring public attitudes of child abuse in mainland China: A sentiment analysis of China's social media Weibo," *Children Youth Services Rev.*, vol. 116, Sep. 2020, Art. no. 105250, doi: 10.1016/j.childyouth.2020.105250.

[47] N. Saleena, "An ensemble classification system for Twitter sentiment analysis," *Proc. Comput. Sci.*, vol. 132, pp. 937–946, Jan. 2018, doi: 10.1016/j.procs.2018.05.109.

[48] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 2931–2939.

[49] P. M. Brennan, J. J. M. Loan, N. Watson, P. M. Bhatt, and P. A. Bodkin, "Pre-operative obesity does not predict poorer symptom control and quality of life after lumbar disc surgery," *Brit. J. Neurosurg.*, vol. 31, no. 6, pp. 682–687, Nov. 2017, doi: 10.1080/02688697.2017.1354122.

[50] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. 1st Int. Conf. Learn. Represent. (ICLR)*, 2013, pp. 1-12.

[51] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Doc2Vec," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 29–30.

**ZIJIAN BAI** received the B.S. and M.S. degrees from the School of Transportation, Jilin University, China, in 2003, and the Ph.D. degree from the Institute of Systems Engineering, Tianjin University, China, in 2007. Since 2007, he has been with Tianjin Municipal Engineering Design and Research Institute Company Ltd., where he is currently the Director and a Chief Engineer of the Urban and Traffic Planning and Design Institute with Management Science and Engineering. His research interests include ITS and urban traffic control and management.

**SHUANGYI MA** received the B.S. degree from the School of Management, Jilin University, Changchun, China, in 2019. She is currently pursuing the M.S. degree with the College of Management and Economics, Tianjin University. Her research interests include natural language processing and machine learning.

**GENG LI** received the B.S. and M.S. degrees from China Agricultural University, Beijing, China, in 2003 and 2007, respectively, and the Ph.D. degree from the College of Management and Economics, Tianjin University, China, in 2016. His research interest includes information security management.