applied sciences

MDPI

Article

A Data-Driven Approach for University Public Opinion Analysis and Its Applications

Miao He [1], Chunyan Ma [2,*] and Rui Wang [3]

[1] School of Marxism, Northwestern Polytechnical University, Xi'an 710129, China
[2] School of Software, Northwestern Polytechnical University, Xi'an 710129, China
[3] School of Automation, Northwestern Polytechnical University, Xi'an 710129, China
* Correspondence: machunyan@nwpu.edu.cn

Abstract: In the era of mobile Internet, college students increasingly tend to express their opinions and views through online social media; furthermore, social media influence the value judgments of college students. Therefore, it is vital to understand and analyze university online public opinion over time. In this paper, we propose a data-driven architecture for analysis of university online public opinion. Weibo, WeChat, Douyin, Zhihu and Toutiao apps are selected as sources for collection of public opinion data. Crawler technology is utilized to automatically obtain user data about target topics to form a database. To avoid the drawbacks of traditional methods, such as sentiment lexicon and machine learning, which rely on a priori knowledge and complex handcrafted features, the Word2Vec tool is used to perform word embedding, the LSTM-CFR model is proposed to realize Chinese word segmentation and a convolutional neural network (CNN) is built to automatically extract implicit features in word vectors, ultimately establishing the nonlinear relationships between implicit features and the sentiment tendency of university public opinion. The experimental results show that the proposed model is more accurate than SVM, RF, NBC and GMM methods, providing valuable information with respect to public opinion management.

Keywords: online public opinion; university public opinion; machine learning; public opinion analysis

check for updates

Citation: He, M.; Ma, C.; Wang, R. A Data-Driven Approach for University Public Opinion Analysis and Its Applications. *Appl. Sci.* **2022**, *12*, 9136. https://doi.org/10.3390/app12189136

Academic Editor: Giacomo Fiumara

Received: 22 August 2022
Accepted: 8 September 2022
Published: 12 September 2022

1. Introduction

Public opinion refers to the views of a majority of people, such as social managers, enterprises, organizations and individuals with respect to the generation, development and impact of social events in the social environment. Public opinion comprises a collection of attitudes, opinions, emotions and other manifestations expressed by the majority of people on various social phenomena and current events. Internet public opinion is generally described as an event that is discussed and considered by many people on the Internet with a considerable social influence, whereby people exhibit tendencies that reflect their psychological state. Internet public opinion is characterized by fast spreading speed, wide spreading routes and considerable social influence [1–4]. With the rapid development of Internet technology, Internet public opinion is an important barometer for public opinion. University online public opinion is a special kind of public opinion that refers to the opinions, views and emotions expressed by college students on social media with respect to academics, life and social issues. On the one hand, college online public opinion reflects students' views and attitudes toward given events and their values; on the other hand, it reflects students' demands and emotional needs. Given that online social media is usually the source of online public opinion, users with smartphones are able to freely publish their opinions. Because young students are immature and intensely curious about a variety of subject, they can be easily affected by one-sided and radical remarks on the Internet. Consequently, they may be consumed by public opinion and even engage in excessive behaviors that lead to serious consequences [5,6].

In comparison to traditional media, the content of online data is updated rapidly, with diverse data forms. Users cannot only send messages and upload pictures but also publish videos and animations on the Internet. Therefore, online public opinion is characterized by a massive volume of data and high levels of uncertainty and complexity. High-value information reflecting sentiments of people is hidden in these unstructured data. However, it is time-consuming and inefficient to manually judge users' attitudes hidden among public opinion, and it is difficult to cope with the ever-growing volume of public opinion data over time. Therefore, effective analysis of university online public opinion has become a popular issue [7].

Content sentiment analysis, also referred to as opinion mining, is the process of making judgments about the sentiments and opinions expressed in text through specific methods and classifying expressions into sentiment categories [8]. Because the remarks posted by students on social media contain exhibit a considerable number of subjective emotional tendencies, they directly influence the direction of online public opinion. Content sentiment analysis provides a technical means to analyze users' psychological needs, which drive online public opinion. According to the granularity of processed texts, sentiment analysis can be divided into word-level, phrase-level, sentence-level and chapter-level analyses. Technically, content sentiment analysis methods mainly include sentiment lexicon, machine learning and deep learning [9]. The sentiment-lexicon-based approach is a simple imitation of human memory and judgment processing, largely relying on sentimental words and judgment rules to classify content sentiment tendencies. Although this method is simple and easy to implement, it is still limited because words have diverse meaning, and sentimental words are absent. With the increasing development of natural language processing technology and computer science, the machine-learning-based content sentiment analysis method has become an important solution for online public opinion analysis [10]. Within the framework of the machine learning method, the sentiment features in the text need to be extracted manually (e.g., using principal component analysis, the information gain method, the frequency feature extraction method, etc.); then, the extracted features are assigned weights (e.g., entropy weights, Boolean weights, etc.), and the text sentiment tendency is derived using classification models (e.g., decision tree, Bayesian classifier, support vector machine, etc.). When appropriate text features are selected and fully trained, machine-learning-based content sentiment analysis outperforms lexicon-based methods. However, machine learning-based methods strongly rely on handcrafted features, mapping of raw Chinese text to abstract features requires extensive human experience and workload, which restricts the further development of this method. As a branch of machine learning, deep learning methods have attracted considerable attention from researchers due to automatic feature engineering capabilities [11]. Using deep networks, raw text data can be represented as progressive, layer-by-layer features, and these features can be learned simultaneously to form an end-to-end content sentiment analysis model, which considerably simplifies the training process.

Motivated by the limitations of sentiment-lexicon-based methods and machine-learning-based methods, in this paper, we propose a deep-learning-based model for analysis of university online public opinion. First, a data crawler is used to automatically acquire data from social media frequently used by college students. Following data preprocessing, word embedding and Chinese word segmentation, raw data can be transformed into numerical word vectors. Finally, a CNN-based deep learning model is used to automatically extract semantic features and output the sentiment tendency of public opinion data. The main contributions of this paper are as follows:

(1) A data crawler is applied to collect public opinion data on social media. In contrast to manual collection methods, data crawlers simultaneously collect data from frequently used Apps to form a high-quality database;

(2) An LSTM network-based Chinese word segmentation model is proposed. Compared with traditional lexicon-based or statistically based word segmentation methods, this method can effectively capture long-term semantic dependencies in sentences and can be self-updated based on new versions of the corpus;

(3) A CNN-based sentiment analysis model is proposed for university public opinion. Compared with traditional machine learning models, the deep learning model does not require tedious feature engineering, and it is capable of automatically extracting high-dimensional features using a deep network structure to form an end-to-end model.

2. Related Works

Online opinion identification is important work. DARPA proposed the TDT (Topic Detection and Tracking) project in 1996 to identify unknown topics in news and track known topics [12]. Simon and Jerit [13] showed that government, networks and media can influence the propagation path of public opinion events, providing theoretical support for public opinion governance. However, public opinion research is usually oriented toward the government, and relatively less attention has been paid to university public opinion.

Research on university online public opinion involves four main aspects: public opinion information collection, topic detection, opinion mining and content tendency analysis. The authors of [14–16] proposed sentimental-lexicon-based classification methods, whereby researchers used statistics indices, including maximum entropy and mutual information, to describe the similarity between the text to be quantitatively detected and the words in the sentiment lexicon and subsequently used manually formulated rules to judge the sentiment tendency. This method is relatively simple and intuitive. However, building a complete sentimental lexicon and rules is too complicated and strongly relies on the designer's a priori knowledge, which is only applicable to some specific fields and difficult to extended to wider ranges.

Machine learning methods can transform text into digital signals that computers can recognize; then, related algorithms can be used to realize content sentiment classification. Pang et al. first used machine learning algorithms for sentiment analysis of movie comments [17]. Results found that the accuracy and recall rates of machine learning methods, such as support vector machine and naive Bayes, on the IMDb dataset outperformed knowledge-based algorithms, verifying the feasibility of machine learning for sentiment classification. David M. Blei proposed a text modelling method based on latent Dirichlet allocation (LDA) [18]. With this method, probability distributions can be used to quantitatively describe the distribution of topics. Accordingly, machine learning classification algorithms can accomplish the task of opinion analysis [19,20]. In ref. [21], public opinion data collected from Baidu Post Bar were treated by the LDA model for text feature extraction, followed by cosine similarity calculation. According to the similarity results, the Krill Herd Harmony search optimization algorithm was used to help a CNN improve classification accuracies. In ref. [5], a machine learning and evolutionary algorithm-based college student public opinion analysis method was proposed, whereby singular value decomposition was applied to perform coordinate transformation within the sample space to determine the abstract meaning of original features. Then, a basic BP-NN was established to perform forward inference after optimizations by the evolutionary algorithm. Inter-information was selected as the text feature classification index for student public opinion management of campus commentary [6].

Although machine-learning-based approaches are usually superior to sentiment-lexicon-based approaches, the disadvantages of machine learning include that training relies on high-quality labelled datasets and appropriate selections of text features (including but not limited to *n*-gram features, TF-IDF features, TFC features, Boolean weights, etc.), which are not easy to handle with a high volume of complex text data. With the rapid development of computer science and big data technology, deep learning has made breakthrough achievements in many fields [22]. Compared with traditional machine learning and sentiment-lexicon-based methods, the advantage of deep learning techniques is that deep networks can automatically extract text features from input data without the need for prior knowledge and manual feature engineering. In ref. [23], a parallel neural network structure was proposed for sentiment classification of MOOC discussion forums. In this structure, features hidden in the discussion texts can be automatically extracted through CNN and LSTM and then for sentiment judgement. Gao, Sun, Wang et al. [24] established a deep learning model for textual information classification, whereby the BiLSTM emotion classification model and the ARIMA timing model were successively employed to treat experimental data. In ref. [25], public opinion data from the Weibo app were analyzed. The authors used an LSTM network to process a structured corpus of text data after preprocessing, then applied a multifeature fusion method to obtain a sentiment trend.

For the purpose of highlighting the motivations of this paper, a summary of articles regarding deep-learning-based and other models for sentimental analysis in the educational field is presented in Table 1. Although a few reports have been published about the important topic of university public opinion analysis, most have focused on the English context, with results that are difficult to generalize to the Chinese context. Moreover, none of the studies listed below simultaneously considered automatic data collection from mainstream social network applications, Chinese word segmentation, automatic feature engineering and a deep learning model for sentimental classification. Therefore, a systematic framework for Chinese data acquisition is required for sentimental analysis.

Table 1. Summary of articles regarding deep learning and other models for sentimental analysis in the educational field.

No.	Reference	Subject	Technique	Feature Extraction	Data Sources
1	[23]	Sentiment classification of MOOC discussion forums	Parallel neural networks	Automatic extraction by CNNs	Stanford MOOC posts dataset (in English)
2	[5]	College student public opinion	Neural network	Singular value decomposition	Offline questionnaire
3	[6]	Student public opinion	Deep learning	Inter-information	Existing campus network public opinion data
4	[24]	Textual information classification of campus network public opinion	BILSTM + ARIMA	Not clearly stated	Standard datasets and a real sample dataset from a paper website
5	[21]	Public opinion analysis in colleges and universities	Intelligent data mining	LDA model	Post bar websites
6	[25]	Campus microblog public opinion	LSTM	Not clearly stated	Weibo

3. Model Construction

In this section, the construction of a data-driven model for university public opinion analysis is described in detail. The overall framework is first introduced, followed by the data collection method, the data pretreatment method, the word-to-vector method, the Chinese word segmentation model and the deep learning model.

3.1. Overall Framework

Considering that deep learning technology can automatically extract effective features from input content through deep networks, in this paper, we propose a deep-learning-based analysis model of university online public opinion for analysis and judgement of the sentimental tendencies contained in text content and subsequent monitoring of university online public opinion.

As shown in Figure 1, the online public opinion analysis model consists of data collection, pretreatment of collected data, a word-to-vector method, Chinese word segmentation, deep-learning model-based sentimental analysis and result output. First, five frequently used apps are selected as the sources of public opinion data. Then, crawlers are used to extract public opinion data from the above-mentioned software to performs. Next, data cleaning and stopping word removal operations are performed on the collected raw data to retain the key information in sentences. Then, LSTM networks are used to adaptively perform Chinese word segmentation, and the segmented words are vectorized to form numerical inputs that the deep learning model can recognize. Using the learning ability of the deep network, the implicit mapping relationship between key features and sentiment tendencies in the input data can be acquired and eventually used for public opinion analysis.

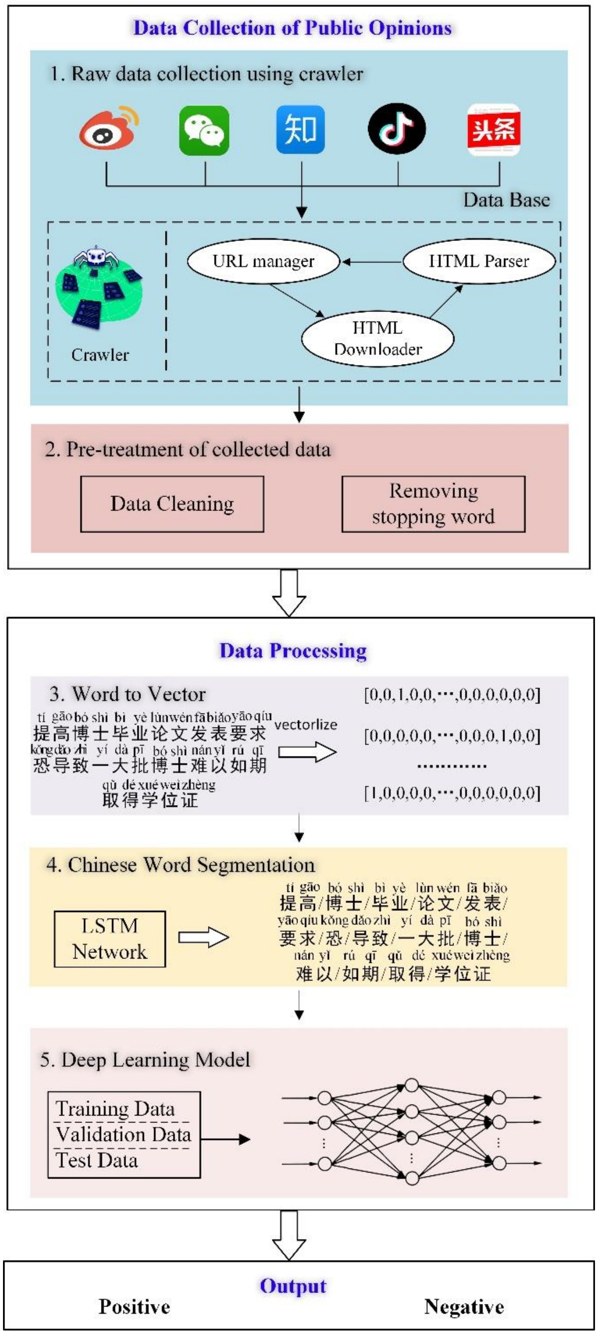


Figure 1. The overall structure of online public opinion analysis model.

3.2. Data Collection

A large amount of data is required to train the proposed public opinion analysis model, whereas the available open dataset is limited. To solve this, five apps frequently used by Chinese university students are selected as data sources, namely Sina Weibo, WeChat, Zhihu, Douyin and Toutiao, as shown in Table 2. Specifically, we select microblog content from Weibo, push content from WeChat, answers from Zhihu, short videos from Douyin and news from Toutiao as the source data; then, a crawler is implemented to automatically extract relevant public opinion data. The crawler used in this study was developed based on the Scrapy framework, and its main workflow includes data acquisition, data parsing, data extraction and data storage. The main modules in this framework are the engine, crawler, scheduler, middleware downloader and project pipeline downloader (as shown in Figure 2).

Table 2. Selected apps and relevant acquisition scopes.

Logo	APP	Scope
	Weibo	Microblog content
	WeChat	Push content
知	Zhihu	Answers
	Tiktok	Short videos
头条	Toutiao	News

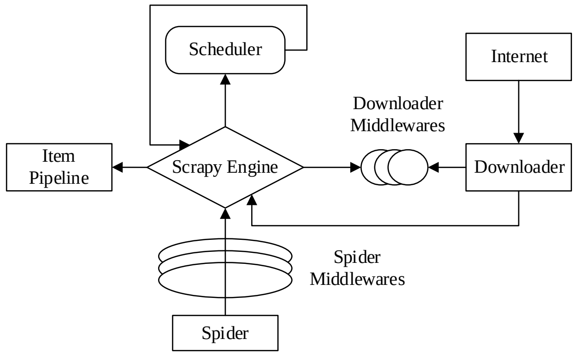


Figure 2. The framework of the crawler.

During the data acquisition process, anticrawler mechanisms, such as HTTP restriction, login restriction, user access restriction, etc., are considered and handled. The collected samples are manually annotated with sentiment attributes for model training and validation.

3.3. Data Pretreatment

URL links, HTML tags, non-Chinese characters and punctuation marks in the raw text have no practical effect on the understanding of public opinion. Moreover, the presence of such information affects the quality of the generated word vectors, and the meaning of such data cannot be accurately expressed. In order to eliminate the abovementioned non-essential information, data cleaning is performed using the regular expressions shown in Table 3 to reduce the redundant information in the original data.

After data cleaning, the collected raw data is stored in plain English and Chinese forms, but these sentences still contain stopping words that are not helpful for sentiment analysis. The stopping words do not contain valid information, and they have negative impacts on the running time, data storage cost and feature space dimension of the algorithm. Therefore, conjunctions, prepositions, tone auxiliaries, modifiers and personal pronouns (see Table 4) are removed to reduce text noise and improve the efficiency and accuracy of the public opinion analysis model.

Table 3. Regular expressions for redundant information removal.

Content	Regular Expression
URL link	((https | http | ftp | rstp | mms)?:W) [^\s]+
HTML label	</?\w+[^>]*>>
Non-Chinese characters	[^\u4e00-\u9fa5]+
Punctuation	[\s+\.\!V_,$%^*(+\″\′]+ | [+——！，。？、~@#￥%……&*（）]+

Table 4. Examples of stopping words.

Part of Speech	Examples
Preposition	zì yú dào wǎng 自、于、到、往
Interjection	me ne ya na ba ma 么、呢、呀、哪、吧、吗
Conjunction	yú shì jiē zhe rán hòu cǐ wai zhì yú 于是、接着、然后、此外、至于
Adjunct Word	de de de 的、地、得
Pronoun	tā tā tā wǒ 他、它、你、我

3.4. Word to Vector

Computers cannot directly process symbolic languages, so they need to be represented numerically and converted into low-dimensional real vectors. Here, the word-embedding operation is accomplished using Word2Vec based on the distributed description method, and the skip-gram model is used for training. Specifically, we set the text window size to 5 and the word vector dimension to 100. The negative sampling technique and Hierarchical Softmax are applied to speed up the training process. The database used for training is the open Chinese word vector corpus provided by researchers at Beijing Normal University and Renmin University of China, which contains repositories from Baidu Encyclopedia, Wikipedia, People's Daily, Zhihu, Sina and other platforms [26]. The pretreated public opinion data are transferred to the pretrained skip-gram model to obtain the corresponding word vectors for subsequent Chinese word segmentation.

3.5. Chinese Word Segmentation

English uses separators to segment words, whereas Chinese sentences are written in a continuous manner without separators between words, so different segmentation methods result in different meanings for a sentence. Consequently, incorrect segmentation combinations result in misunderstanding of the original sentence and affect the accuracy of sentiment analysis. Therefore, the key to Chinese word segmentation is reasonably slicing the strings so as to correctly identify words. From a technical perspective, the Chinese word segmentation task can be regarded as a character-based sequence-labelling problem, whereby each character is labelled as one of four categories: beginning (B), middle (M), end (E) or single segmentation (S). Conventional Chinese word segmentation methods include lexicon-based segmentation methods (e.g., maximum matching method) and word-frequency-based segmentation methods (e.g., n-gram, MaxEnt, CRF, etc.); however, the effect of the former is easily affected by the quality of the lexicon, and the implementation of the latter requires a high computational payload. In recent years, neural-network-based word segmentation methods have been widely applied and achieved positive results.

In this study, a long short-term memory network-based Chinese word segmentation model is designed, as illustrated in Figure 3. Considering that the concept of words only applies after word segmentation, a character vector is obtained by the Word2Vec tool, as described in the Section 3.4. For a given text sequence ($\boldsymbol{c}^{(1:n)}$) of length n, a window of size k is slid from the first character ($c^{(1)}$) to the last character ($c^{(n)}$) (k = 3 is demonstrated in

Figure 3). Then, the vector ($c^{(\mathrm{t}-1)}$, $c^{(\mathrm{t})}$, $c^{(\mathrm{t}+1)}$) is input to the Word2Vec model to obtain the numerical word vectors as the input for the hidden layer. The hidden layer uses a long short-term memory network to extract the features of the input data. In comparison with the RNN model, the LSTM model can better capture the long-term dependencies among sequence elements and achieves better training results [27]. The LSTM unit consists of an input gate, a forget gate and an output gate. The mathematical updating process is shown in Equations (1)–(6) [27], where $i_\mathrm{t}, f_\mathrm{t}$ and o_t are the input gate, forget gate and output gate, respectively; and $\sigma(\cdot)$ represents the sigmoid function.

$$i_t = \sigma(\boldsymbol{W}_i[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + b_i) \tag{1}$$

$$\widetilde{C}_t = \tanh(\boldsymbol{W}_c[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + b_c) \tag{2}$$

$$f_t = \sigma\left(\boldsymbol{W}_f[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + b_f\right) \tag{3}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \widetilde{C}_{t-1} \tag{4}$$

$$o_t = \sigma(\boldsymbol{W}_o[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + b_o) \tag{5}$$

$$h_t = o_t \circ \tanh(C_t) \tag{6}$$

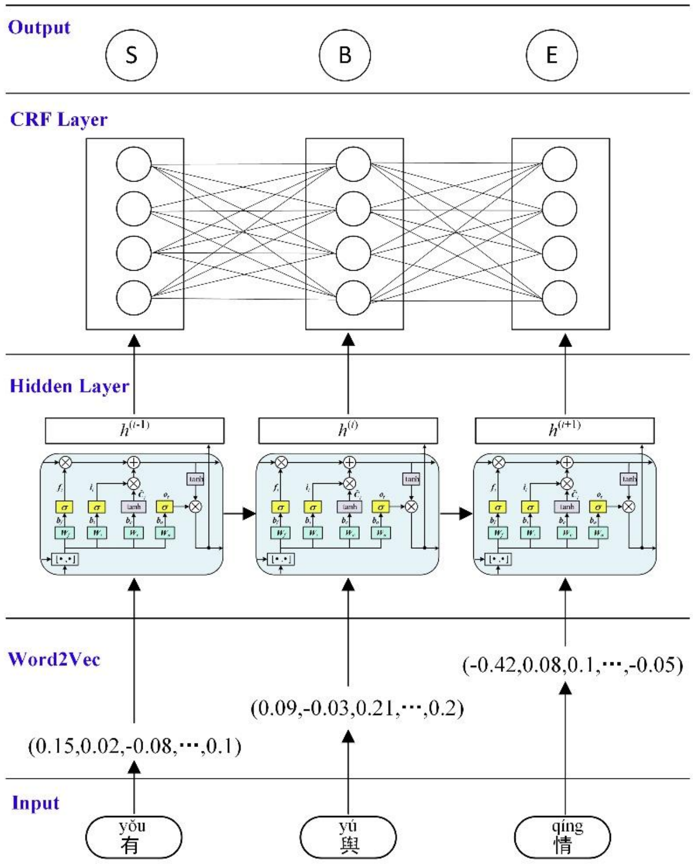


Figure 3. Chinese word segmentation architecture.

The features extracted from the LSTM layer are input to the CRF layer for four-tag labelling. In the CRF layer, the input sequence is first scored according to Equation (7), where y_i represents the label value, h_i represents the sequence output value of the LSTM network and matrix $\boldsymbol{A}$ is the model parameter.

$$S(X, y, \theta) = \sum_{i}^{N} A_{y_i, y_{i+1}} + \sum_{i}^{N} h_{i, y_i} \tag{7}$$

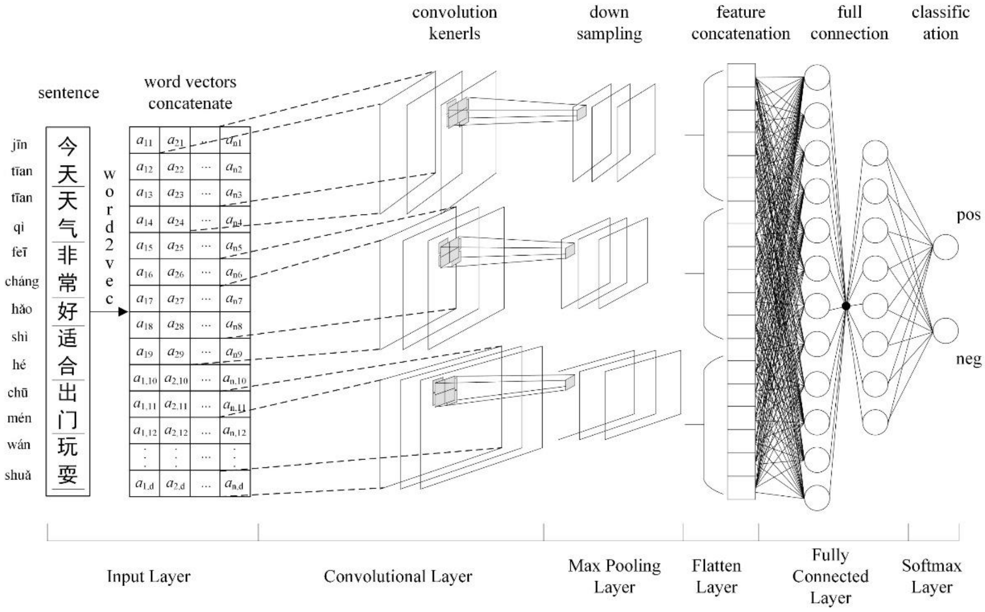
Then, the CRF layer defines chained conditional probabilities to normalize the output scores, as shown in Equation (8).

$$P(y|x) = \frac{\exp(S(X, y, \theta))}{\sum_{y_i} \exp(S(X, y_i, \theta))} \tag{8}$$

Accordingly, Chinese word segmentation results of the input sequence can be obtained and utilized for sentiment analysis through the deep learning model.

3.6. Deep Learning Model

The deep learning model is the key part of the proposed university public opinion analysis model, and it is introduced in this subsection. Because the structure of convolutional neural networks is based on local perception and weight sharing, high-dimensional semantic features can be extracted through the deep neural network to better understand the sentiment tendency. The architecture of the deep learning model is illustrated in Figure 4. The input layer uses a pretrained Word2Vec model to map the segmented Chinese text (containing *n* words) into the word vector space (word vector dimension, *d*), and these word vectors are merged to form a two-dimensional matrix as the input of the convolutional layer. The convolutional kernel is the core component of the convolutional neural network. By performing convolution operations in the input matrix, corresponding high-level feature maps can be established. In order to obtain more content features, three convolution kernels with dimensions of 3 × *d*, 4 × *d* and 5 × *d* (where *d* is the word vector dimension) are used in the model. Given the large size of the feature map, a pooling operation is used to realize the downsampling of feature maps to reduce the dimensionality and decrease the computational effort. Maximum pooling is selected in the model, and the pooling window size is set to 2 × 2, whereas the slide step is set to 2. Then, the size of feature maps is effectively reduced, and the three feature maps are concatenated and flattened to form a one-dimensional vector. Two fully connected layers are connected after the flattened layer to map the semantic features acquired by the convolutional network to sentiment labels. A Softmax classifier calculates the probability that the feature vector belongs to positive or negative comments and determines the sentiment tendency of the text data based on the magnitude of the probability value. In the deep learning model, the Relu function is used as the activation function to accomplish gradient backpropagation.



a_{11}	a_{21}	…	a_{n1}
a_{12}	a_{22}	…	a_{n2}
a_{13}	a_{23}	…	a_{n3}
a_{14}	a_{24}	…	a_{n4}
a_{15}	a_{25}	…	a_{n5}
a_{16}	a_{26}	…	a_{n6}
a_{17}	a_{27}	…	a_{n7}
a_{18}	a_{28}	…	a_{n8}
a_{19}	a_{29}	…	a_{n9}
$a_{1,10}$	$a_{2,10}$	…	$a_{n,10}$
$a_{1,11}$	$a_{2,11}$	…	$a_{n,11}$
$a_{1,12}$	$a_{2,12}$	…	$a_{n,12}$
⋮	⋮	…	⋮
$a_{1,d}$	$a_{2,d}$	…	$a_{n,d}$

Figure 4. Architecture of the deep learning model.

4. Results and Discussion

4.1. Experimental Data

A crawler is employed in the aforementioned social media platforms with the keywords "university, doctoral students, postponed graduation, thesis publication" during the period from 1 January 2022 to 30 June 2022. A total of 74,330 samples are obtained, including 21,556 Weibo samples, 18,583 WeChat samples, 15,609 Douyin samples, 7433 Zhihu samples and 11,149 Toutiao samples. Data cleaning and stopping word removing are applied to the original samples according to the method described in subsection "Data Pretreatment". The ratio of data set to test set to validation is set to 8:1:1, as shown in Table 5. The positive sample shows support or consent in public opinion, such as supporting the abolition of strict restrictions on publication by doctoral students, appropriate relaxation of journal requirements for the publication of papers and increased emphasis on the views presented in papers rather than authorship. The negative sample expressed opposition or denial of the petitioners in matters of public opinion, such as questioning why petitioners should seek special treatment to lower standards, as others can complete work on time. These samples are used for model training.

Table 5. Distribution of the dataset.

Data Set	Positive Samples	Negative Samples	Total Samples
Training set	26,623	32,843	59,466
Validation set	3327	4105	7432
Test set	3327	4105	7432

4.2. Evaluation Indicators

In order to quantitatively describe the effectiveness of the model, precision, recall and F_1 values are used to evaluate the sentiment classification performance. Precision refers to the ratio of the number of correctly classified samples in a category to the total number of samples in that category; recall refers to the ratio of the number of correctly classified samples in a category to the number of samples that actually belong to that category; and F_1 value is the weighted average value of precision and recall, which is a comprehensive index to judge the performance of the classifier. The precision rate, recall rate and F_1 value are calculated as follows:

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{11}$$

where *TP*, *FP*, *FN* and *TN* are defined in Table 6.

Table 6. Definition of *TP*, *FP*, *FN* and *TN*.

Actual Prediction	Positive	Negative
Positive	*TP*	*FP*
Negative	*FN*	*TN*

Based on the above indicators, an accuracy indicator is proposed to represent the classification performance of the model.

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \tag{12}$$

4.3. Implementation Details

The hardware platform and software platform used in the experiment are listed in Table 7. In the experiment, the pretraining word vector database (CA8 Chinese word vector corpus) is first loaded and read using the Gensim library [26], the Word2Vec skip-gram model is pretrained using these data with negative sampling (sample size of 20), and the hierarchical Softmax technique is applied. With respect to training, the initial learning rate is set to 0.01, the size of the sliding window is 5 and the lowest word frequency is determined as 3. Then, the Bakeoff 2005 corpus is used as the Chinese word segmentation training dataset [28], and simplified Chinese corpora, including Peking University (PKU) and Microsoft Research Associates (MSRA) datasets, are ultimately selected. In this paper, the 4-g annotation method is used for word segmentation, whereas the training corpus does not provide the required labels (corpus is segmented only by spaces). Therefore, the corpus data are manually labelled before training, and each Chinese word in each sentence is labelled as either B, M, E or S. Considering that there is no concept of a word before Chinese word segmentation, the pretreated data are sent to the Word2Vec model character by character to generate character vectors. Based on these vectors, the Chinese word segmentation model provides appropriate segmentation results, and segmented words are again sent to the Word2Vec model to produce corresponding word vectors. The deep learning model takes the word vectors as input for the final sentiment classification. The implementation procedure is shown in Figure 5.

Table 7. Experimental details of the hardware and software platforms.

Experimental Setup	Details
OS	Windows 10 64 bit
CPU	Intel Xeon W2223
Memory	16 GB
GPU	P2200
Language	Python
IDE	Pycharm
Major packages	Requests, Gensim and Pytorch

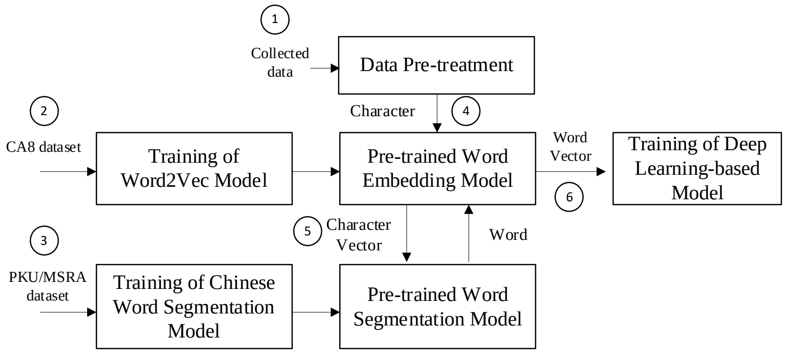

1
Collected data
Data Pre-treatment
Character
4
2
CA8 dataset
Training of Word2Vec Model
Pre-trained Word Embedding Model
Word Vector
6
Training of Deep Learning-based Model
3
PKU/MSRA dataset
5
Character Vector
Word
Training of Chinese Word Segmentation Model
Pre-trained Word Segmentation Model

Figure 5. Training procedure of the proposed model.

In the Chinese word segmentation model, the dimension of the LSTM unit is 200, the batch size is 32, the optimizer is Adam and the dropout technique with a ratio of 0.2 is used to prevent overfitting. In the online public opinion analysis model, sizes of convolutional kernels are set to 3 × 100, 4 × 100 and 5 × 100, and the number of three-convolutional kernels is 80. The number of nodes in each of the two fully connected layers is 100 and 60, respectively, and the output layer uses Softmax as the classifier. To prevent overfitting, the dropout technique is also introduced in the deep learning model with a ratio of 0.5. Given the relatively high volume of the collected opinion data, the batch size is set to 128, and the maximum number of training iterations is 32. RMSprop is selected as the optimizer, and the learning rate is set to 10^{-3}.

4.4. Model Performance

Chinese word segmentation is the input of the public opinion analysis model; thus, the accuracy of Chinese word segmentation directly determines the accuracy of sentiment classification. Figure 6 shows the results of the proposed Chinese word segmentation model on the MSRA and PKU datasets. The F_1 value increases as the number of iterations increases, reaching 95.8% and 94.2% on each of the two datasets, respectively.

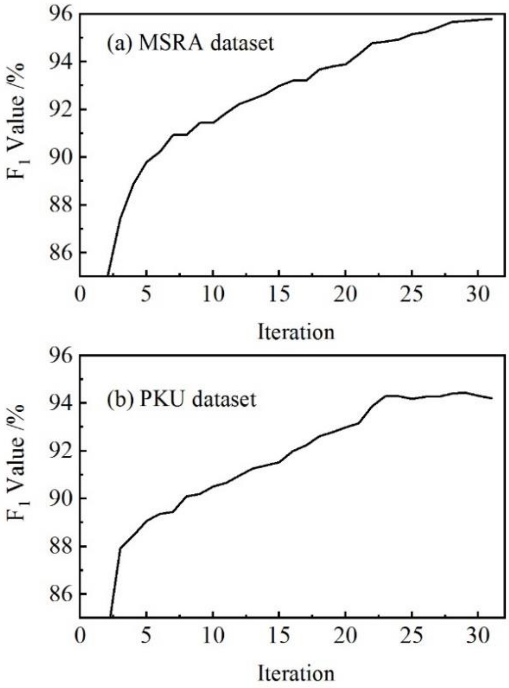


Figure 6. F_1 values of the Chinese segmentation model.

Table 8 shows the accuracy, recall and F_1 value of the proposed word segmentation model in comparison with the LTP, NLPIR, Thulac and Jieba methods on the MSRA and PKU datasets. Experimental results suggest that the LSTM-CRF model with Word2Vec pretreatment achieves the best word segmentation effect. The Word2Vec model can quantitatively describe the similarity between words and establish mapping between Chinese words and word vectors. Based on this relationship, the LSTM unit can effectively capture the dependencies between sequential texts, grasp the abstract features in the text and ultimately map the features to segmentation sites through the CRF layer.

Table 8. Performance comparison of the Chinese word segmentation model.

	MSRA			**PKU**		
	P	***R***	***F***	***P***	***R***	***F***
Proposed	0.962	0.955	0.958	0.944	0.940	0.942
LTP	0.865	0.893	0.835	0.916	0.895	0.911
NLPIR	0.866	0.912	0.895	0.936	0.941	0.937
Thulac	0.931	0.922	0.925	0.943	0.911	0.921
Jieba	0.815	0.811	0.819	0.853	0.799	0.821

Table 9 shows the performance of the deep learning model, in which binary logistic regression (BLR), random forest (RF), support vector machine (SVM), naive Bayes classifier (NBC) and Gaussian mixture models (GMM) are used to compare with the proposed method. Because BLR, RF, SVM, NBC and GMM are machine-learning-based approaches, their implementation requires handcrafted features. The effect of machine learning models considerably depends on feature selection, and they do not perform well in the experiments. The deep learning model can automatically extract deep semantic features in sentences on the basis of word vectors, and the semantic features are high-dimensional features that can better reflect the implicit information in the opinion corpus. Without complex feature engineering, this end-to-end model achieves excellent classification performance, with accuracy, recall and F_1 values of 78.2%, 74.5% and 76.7%, respectively. To further validate the classification performance, the model is evaluated on open datasets, including IMDb, SST-2 and MRPC. As shown in Figure 7, the established deep learning model achieves better results on the standard datasets than the self-built dataset, with classification accuracies of 88.5%, 87.6% and 85.1% achieved on the IMDb, SST-2 and MRCA datasets, respectively, also outperforming popular handcrafted-feature-based machine learning methods, such as SVM and NBC. Based on these results, the proposed deep-learning-based public opinion analysis model is effective in classifying sentimental tendencies hidden in Chinese text content.

Table 9. Performance comparison of public online analysis models.

Model	*P*	*R*	F_1
Proposed	0.782	0.745	0.767
Binomial logistic regression	0.484	0.406	0.459
Random forest	0.650	0.712	0.695
SVM	0.701	0.731	0.715
Naïve Bayes classifier	0.562	0.495	0.528
GMMs	0.576	0.547	0.561

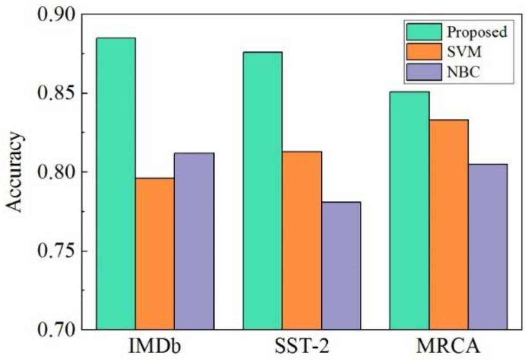


Figure 7. Model performance on open datasets.

4.5. Further Discussion

Based on the collected data and the tendency results, we can further analyze the public opinion on the topics of "universities, doctoral students, delayed graduation and thesis publication".

Figure 8 illustrates the tendency distribution of positive and negative public opinions on various social media platforms. Although the general trend is negative, the distribution tendency varies depending on the due to the difference in interactivity. For example, Douyin and WeChat are strongly personalized social network, and the identities of users are easy to recognize; thus, sentiment trends in these apps are positive. As a representative platform for short videos, the ratio of positive attitudes as high as 71%, with the following main opinions represented: "Strict management should be imposed on Ph.d"; "There are reasons for both supervisors and students to postpone graduation"; and "Study should be the first priority". Internet users who are concerned about such topics are influenced by scenario-based immersion. Such topics are related to higher education and professionalism, and positive opinions show that users have a high degree of understanding of universities and the social need to take a positive stand. In contrast, negative attitudes of users are more prominent on platforms such as Toutiao, Weibo and Zhihu, wherein interaction is open but identities is hidden. The main opinions expressed on these platforms include: "Tutors squeeze students in various ways"; "Students are free workers"; "Student cadres should not be directly recommended to be postgraduate candidates", etc. Among the Toutiao, Weibo and Zhihu apps, the highest ratio of negative trends is 77% (Toutiao), possibly as a result of the one-way output nature of these apps. Although Weibo seems to be interactive, information is published in a free market, and the polarization phenomenon is prominent, reflecting that users are influenced by online public opinion, resulting in emotional expressions. Zhihu prominently features negative evaluation attitudes, representing a critical thinking style.

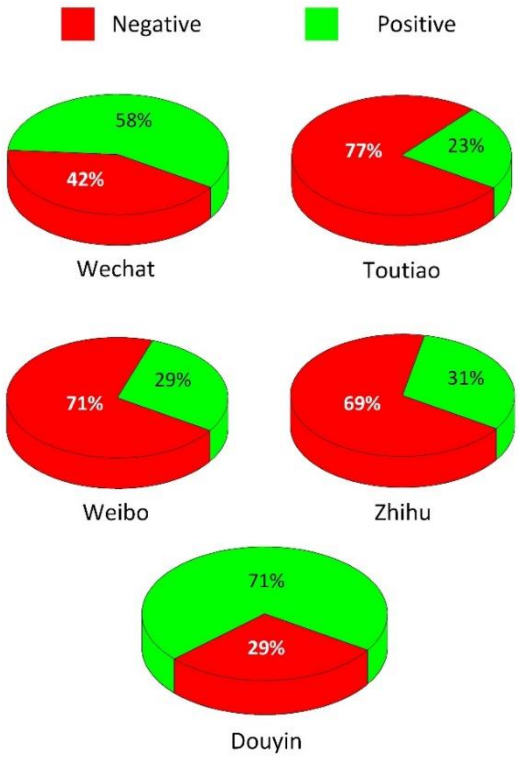


Figure 8. Sentiment trends in social media apps.

Figure 9 shows the extent to which users on Weibo have moved from wait-and-see to action attitudes on this topic. The peak of behavior difference fluctuates with the evolution of public opinion, including original posts, repost, and repost-and-comment behavior. In the investigated dataset, the volume of public opinion peaked in late March, when a participation value of 3264 was generated on Weibo in one day, mainly comprising repost-and-comment behavior, indicating a high participation rate among users on Weibo and the low sensitivity of the topic. After a period of evolution of public opinion, repost-and-comment behavior decreased to a value of 3057 a day before dissipating. Repost-and-comment behavior is the main choice of users on Weibo according to the investigated dataset.

Figure 10 shows a keyword cloud (translated from Chinese words) of public opinion data; "Zhejiang University", "examinations", "postpone", "dissertation", "check", "knowledge network", "university ranking", "dissertation writing" and "core journals" are keywords with high relevance, reflecting the focus of public attention.

The proposed data-driven approach can effectively classify the sentimental tendencies hidden in social networks. In addition, the theoretical implication of the proposed approach is that a deep-learning-based systematic framework including data collection, data pretreatment, Chinese word segmentation, feature extraction and sentimental classification is built to address the university public opinion analysis problem, which is novel approach to this issue. From the perspective of practice, this model could be applied to analysis other current topics and help university administrators identify actual views of students and to inform decisions with respect to emergency events.

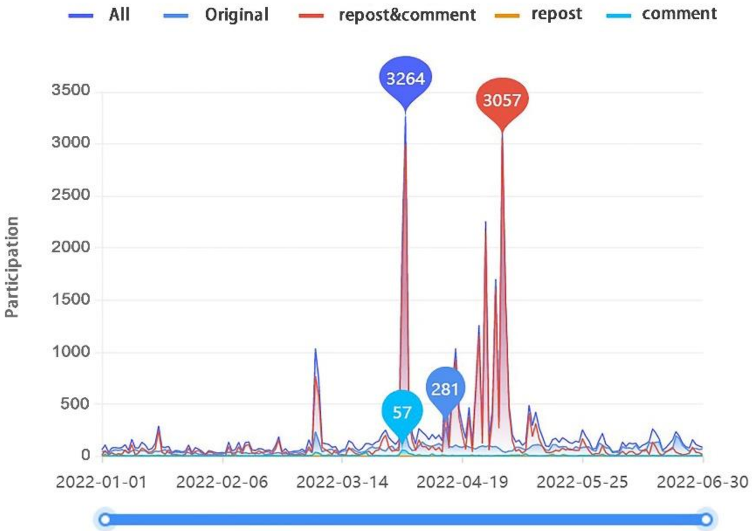


Figure 9. Evolution of public opinion on Weibo.

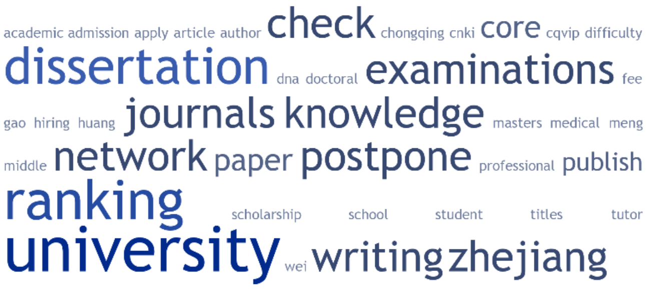


Figure 10. Keyword cloud (translated from Chinese).

5. Conclusions

In this paper, a deep-learning-based systematic framework is proposed for analysis university online public opinion. This framework specializes in the sentimental analysis of current events related to the university environment. Within the proposed framework abstract features hidden in the Chinese sentences are automatically extracted without handcrafted feature engineering. In this end-to-end framework, crawler technology quickly obtains the latest opinion data from the target app and provides textual input for opinion analysis. The word embedding and Chinese word segmentation models map Chinese text to vector space and form a word vector with statistical meanings. The deep learning model takes the word vector matrix as input and automatically extracts the high-level features of the word vectors through deep networks, establishing a nonlinear correlation between the hidden features and the tendencies of university online public opinion. The experimental results show that the established public opinion analysis model can effectively judge the sentiment tendencies of public opinion and could be used to helping university administrators identify actual views of students and inform decisions with respect to emergency events.

However, a major limitation of this framework is that precise classification relies on abundant sentimental samples with manual labels, which may be inconvenient in some cases. In future studies, the structure of the proposed deep learning model should be updated to achieve satisfactory results, even when the volume of the dataset is low.

Author Contributions: Conceptualization, investigation, resources, writing, M.H.; methodology, C.M.; software, R.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was sponsored in part by the Shaanxi Provincial Social Science Fund (2020M008) and Fundamental Research Funds for the Central Universities (GJJJZ202203).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare that they no conflicts of interest regarding this work.

References

1. Burstein, P. The impact of public opinion on public policy: A review and an agenda. *Political Res. Q.* **2003**, *56*, 29–43. [CrossRef]
2. Lippmann, W.; Curtis, M. *Public Opinion*; Routledge: London, UK, 2017.
3. McGregor, S.C. Social media as public opinion: How journalists use social media to represent public opinion. *Journalism* **2019**, *20*, 1070–1086. [CrossRef]
4. Bilal, M.; Gani, A.; Lali, M.I.U.; Marjani, M.; Malik, N. Social profiling: A review, taxonomy, and challenges. *Cyberpsychology Behav. Soc. Netw.* **2019**, *22*, 433–450. [CrossRef] [PubMed]
5. Zhang, J.; Zhang, P.; Xu, B. Analysis of college students' public opinion based on machine learning and evolutionary algorithm. *Complexity* **2019**, *2019*, 1712569. [CrossRef]
6. Shen, L.; Xu, M. Student Public Opinion Management in Campus Commentary Based on Deep Learning. *Wirel. Commun.-Tions Mob. Comput.* **2022**, *2022*, 2130391. [CrossRef]
7. Dong, X.; Lian, Y. A review of social media-based public opinion analyses: Challenges and recommendations. *Technol. Soc.* **2021**, *67*, 101724. [CrossRef]
8. Hemmatian, F.; Sohrabi, M.K. A survey on classification techniques for opinion mining and sentiment analysis. *Artif. Intell. Rev.* **2019**, *52*, 1495–1545. [CrossRef]
9. Li, Z.; Fan, Y.; Jiang, B.; Lei, T.; Liu, W. A survey on sentiment analysis and opinion mining for social multimedia. *Multimed. Tools Appl.* **2019**, *78*, 6939–6967. [CrossRef]
10. Hasan, A.; Moin, S.; Karim, A.; Shamshirband, S. Machine learning-based sentiment analysis for twitter accounts. *Math. Comput. Appl.* **2018**, *23*, 11. [CrossRef]
11. Wang, S.; Cao, J.; Yu, P. Deep learning for spatio-temporal data mining: A survey. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 3681–3700. [CrossRef]
12. Allan, J.; Harding, S.; Fisher, D.; Bolivar, A.; Guzman-Lara, S.; Amstutz, P. Taking topic detection from evaluation to practice. In Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, 3–6 January 2005.
13. Simon, A.F.; Jerit, J. Toward a theory relating political discourse, media, and public opinion. *J. Commun.* **2007**, *57*, 254–271. [CrossRef]
14. Turney, P.D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *arXiv* **2002**, arXiv:cs/0212032.
15. Hu, M.; Liu, B. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 22–25 August 2004.
16. Ding, X.; Liu, B.; Yu, P.S. A holistic lexicon-based approach to opinion mining. In Proceedings of the 2008 International Conference on Web Search and Data Mining, New York, NY, USA, 11–12 February 2008.
17. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 6 July 2002.
18. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
19. Zhuang, M.; Li, Y.; Tan, X.; Xing, L.; Lu, X. Analysis of public opinion evolution of COVID-19 based on LDA-ARMA hybrid model. *Complex Intell. Syst.* **2021**, *7*, 3165–3178. [CrossRef] [PubMed]
20. Ni, N.; Guo, C.; Zeng, Z. Public Opinion Clustering for Hot Event Based on BR-LDA Model. In *International Conference on Intelligent Information Processing*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
21. Wu, W.N.; Deng, Z.H. The Analysis of Public Opinion in Colleges and Universities Oriented to Wireless Networks under the Application of Intelligent Data Mining. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 7597366. [CrossRef]
22. Shinde, P.P.; Shah, S. A review of machine learning and deep learning applications. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 16–18 August 2018.
23. Gao, Y.; Sun, X.; Wang, X.; Guo, S.; Feng, J. A parallel neural network structure for sentiment classification of MOOCs discussion forums. *J. Intell. Fuzzy Syst.* **2020**, *38*, 4915–4927. [CrossRef]
24. Wang, W. Textual Information Classification of Campus Network Public Opinion Based on BILSTM and ARIMA. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 8323083. [CrossRef]

25. Lv, Z. Prediction of the Forwarding Volume of Campus Microblog Public Opinion Emergencies Using Neural Network. *Mob. Inf. Syst.* **2022**, *2022*, 3064266. [CrossRef]
26. Li, S.; Zhao, Z.; Hu, R.; Li, W.; Liu, T.; Du, X. Analogical reasoning on chinese morphological and semantic relations. *arXiv* **2018**, arXiv:1805.06504.
27. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
28. Emerson, T. The second international Chinese word segmentation bakeoff. In Proceedings of the fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea, 14–15 October 2005.