Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 18 (3): 2632–2653.
DOI: 10.3934/mbe.2021134
Received: 06 January 2021
Accepted: 03 March 2021
Published: 19 March 2021

Research article

Research on knowledge dissemination in smart cities environment based on intelligent analysis algorithms: a case study on online platform

Chengzhi Jiang[1,2], Hao Xu[1,2]*, Chuanfeng Huang[2], Yiyang Chen[3], Ruoqi Zou[4] and Yixiu Wang[4]

[1] School of Information Management, Nanjing University, Nanjing 210023, China
[2] School of Economics and Management, Nanjing Institute of Technology, Nanjing 211167, China
[3] School of Computer Science, University of St Andrews, St Andrews KY16 9AJ, United Kingdom
[4] Ping An International Smart City Technology Co., Ltd, Shenzhen 518002, China

* **Correspondence:** Email: xhnju2014@163.com; Tel: +8618151005729.

Abstract: In developing smart cities, the implementation of social connections, collaboration, innovation, exchange of views by observing, exploiting and integrating various types of knowledge is required. The smart cities concept that employs knowledge sharing mechanism can be defined as the concept of a city that utilizes information technology to increase citizens' awareness, intelligence as well as community's participation. The knowledge dissemination via online sharing platforms has been becoming more popular in recent years, especially during the epidemic of infectious diseases. Thus, the social network and emotional analysis method based on intelligent data analysis algorithms is proposed to study the speaker relationship and comment sentiment tendency of a Chinese popular speech (knowledge dissemination) platform: YiXi. In our research, 690 speakers' information and 23,685 comments' information are collected from YiXi website as the data source. The speaker relationship network construction algorithm and emotional analysis algorithm are designed in details respectively. Experiments show that speakers who have the same profession can deliver different types of speeches, indicating that selection of YiXi platform in the invitation of speakers is diversified. In addition, overall sentiment tendency of comments on speeches seem to be slightly positive and most of them are the personal feelings according to their experience after watching speech videos instead of the direct evaluations of speech quality. The research aims to gain an insight into the popular knowledge sharing phenomenon and is expected to provide reference for knowledge dissemination platforms in order to improve the knowledge sharing environment in smart cities.

Keywords: speech; relationship network; sentiment analysis; textual analysis; smart cities

1. Introduction

The majority of cities in the world are trying to improve their cities to achieve smart environment [1]. The governors of smart cities have been focusing on enhancing the 'smart' aspects of their cities in order to improve the quality of life for their citizens, both in material word and mental world. The knowledge network has been identified as an important role that leverages knowledge resources to improve smart cities environment [2]. Nowadays, smart city citizens use social networks to acquire information, share their feelings, exchange views with each other, etc. [3]. Hence, the online platform and community may provide the media for knowledge sharing network. With the help of Internet technology, the depth and breadth of knowledge dissemination have been continuously expanded in recent years, and the development of new media knowledge sharing platform has become increasingly prosperous. From "Ted talks" in United States to "Ge Zhi Lun Dao" and YiXi in China, they have been widely concerned by the public. In many Chinese new media knowledge platforms, YiXi as an independent media combining theater style live speech and network video transmission, carries out content dissemination through live speech combined with online video. Since its establishment in 2012, due to its rapid development in the form of new media communication, flexible mode, rich and distinctive speech content and diverse carriers, nearly 700 speakers from different professional fields with rich research and practical experience have stood on the stage of YiXi to share domain knowledge and promote knowledge dissemination. The most popular speech in YiXi has been watched by over 9 million people, demonstrating the degree of attention paid by the public.

The positioning of YiXi aims to share humanities, technology, and daydreams, and to encourage the sharing of opinions, experiences and future phenomena. Therefore, it has strong inclusiveness in the direction of speech, accommodating a variety of speakers and paying special attention to the richness and inclusiveness of speakers. The social recognition and dissemination of YiXi speech show that the speeches given by selected speakers are considered as good stories. The popular phenomenon of YiXi motivates us to explore the relationship between the careers of diverse speakers and the topics of their speeches first of all. Is there a specific network relationship between them? If their professional characteristics have an obvious impact on the topic selection of their speeches? Meanwhile, in the process of knowledge dissemination, YiXi pays attention to the advantages of new media, especially the digital and interactive characteristics of new media. In the production process of speech content, YiXi actively strengthens the communication with the audience through official website, Weibo and App, and the online audience also expresses their attitude or views on the content of YiXi in the form of comments. Hence, it also motivates us to find out whether the online review texts of YiXi video have obvious emotional tendencies or categories. Could the analysis of comments help platform operators invite better "story" narrators, or further promote the knowledge sharing environment in smart cities?

Thus, this paper takes speaker relationship and comment sentiment analysis of YiXi as the research content, and uses the social network analysis method and text mining method of intelligent analysis method to carry out the research from the data sources of the speakers' information and audiences' comments on YiXi. It tries to explore the mechanism behind the phenomenon, providing a reference for improving knowledge dissemination in smart cities environment.

2. Related works

Researches on big data based smart city applications have been attracting scholars. Cai et al. used data collected from bus Automatic Fare Collection system to train and test their passenger estimate model, aiming to improve bus transit service in smart cities [4]. Sun et al. proposed a clustering method to analyze people flow in smart cities environment and the method was applied to real taxi dataset of a Chinese city [5]. In addition, some image data process algorithms were proposed and tested on public data sets [6–8].

Among the relationship identification methods, social network analysis method is a classic method. Its application fields cover natural science and social science research fields. This method is a set of norms and methods to analyze the structure and attributes of social relations. It mainly analyzes the relationship structure and attribute relationship formed by different social units [9]. In the field of natural science, the main goal of social network analysis method is to reveal the comprehensive characteristics of non- random networks, and to pay attention to whether some representative networks have unique properties different from other networks, such as the phenomenon of rich club in scientists' cooperative networks [10]. In sociological research, researchers tend to study the structural changes among different organizations, and explain the reasons for their inconsistent output from the perspective of sociology. However, whether in the field of natural science or social science, their common research goal is to explain the formation mechanism of network links and predict the characteristics of the network [11]. In terms of social network algorithm, the main algorithms involved include GN algorithm, Newman greedy algorithm, Louvain algorithm and SLM algorithm [12–15]. In view of the breadth and applicability of social network analysis, this study is based on the relationship between speaker occupation and topic selection.

Text mining is a relatively new research field in the field of big data analysis. Related research mainly involves five categories: emotion and topic analysis, concept and semantic relationship discovery, biomedical research, text mining theory and main algorithm model, and other applications. Some studies also show that it has been used in the fields of opinion mining, document classification, document core content mining and so on. Erhan et al. use the smartphone app to collect the subjective data of participating citizens including comments about what they noticed, why they are in that place, etc. The text analysis of their research categorizes comments to 11 classes that gave an idea of the kind of activities people were engaging in [16]. Xu et al. proposed a recommendation method to capture textual matching of users and items via review network for smart city applications [17]. Kilicay-Ergin and Barb explored a method to extract domain knowledge from text and applied it to the smart city domain, looking for the relevance of news articles to the topic of smart city [18]. Oza and Naik takes the review data in the online learning process as the basic data set, analyzes the data through the text clustering process in text mining, and judges the popularity of online courses by identifying the generated topic words [19]. Based on the review data of search products, Ghose and Ipeirotis analyzes the subjective and objective tendency of reviews, and evaluates the impact of online review usability with the help of subjective and objective tendency mixing index [20]. By establishing a user evaluation usefulness model, Susan and David analyzes the impact of the extreme nature of the comment text, the depth of the comment and the commodity type on the perceived comment usefulness [21]. Based on text mining technology and empirical research methods, Chen et al. explored the relevant influencing factors of the usefulness of text-based reviews, aiming at the problems of the overall quality of online product reviews and the lack of effective comment guidance mechanism [22]. Based on the theory of information adoption, Yin et al. discussed two kinds of influencing factors of consumers' adoption and acceptance of online review information in purchase decision-making, that is, the characteristics of reviews and the elements of reviewers, and constructed an online review usefulness influence model from the perspective of social network [23]. Founoun et al. reviewed the regulations related to smart cities using textual analysis of similarities between local environmental regulations to key factors in smart environment [24]. Yu et al. used text mining technology to analyze Chinese Tang Dynasty poetries so as to automatically generate Chinese poetry [25].

From the perspective of review information itself and its related factors, the above research identified that user evaluation will have an important impact on user behavior. Therefore, it is of great application value to take the comment data as the "preference degree" of users for something as an important component of recommendation strategy, and then carry out the recommendation of related products and identification of user stickiness.

This research focuses on text sentiment analysis, which is essentially a kind of research to analyze people's views, emotions, evaluations, attitudes and emotions from written language with the help of natural language processing, text mining and computer linguistics. At present, the commonly used text sentiment analysis methods are mainly based on emotion dictionary and machine learning. There are many researches on sentiment analysis of user comments based on text mining technology. For example, Zhang classifies the comment text of a brand water heater in Jingdong e-commerce platform into three categories: positive, negative and neutral through the emotional analysis in ROSTCM6, and eliminates the neutral emotional text [26]. By constructing LDA theme model, the "positive emotional results" and "negative emotional results" are analyzed, and the theme set and keywords of the review text are obtained, reflecting the advantages and disadvantages of the product in the market competition. Li and Yu established an emotional analysis model for film and television products [27]. After preprocessing the collected TV drama information, such as text de-duplication, mechanical compression and word segmentation, LDA model and relational network were applied to model and analyze the themes and comments, and the audience's concerns, subjective evaluation and feelings of the film and television works were mined out, and the practicability of the model was proved. Wu et al. applied text mining technology to the study of spatial human geography [28]. By analyzing the characteristics, advantages and scale characteristics of ancient poetry texts, they selected Tang poetry for text mining. From the macro to micro level, they explored the emotional characteristics, spatial semantics and human landscape pattern of Guanzhong area, summarizing the portrait and image characteristics to obtain the relationship of color, element and space of the humanistic landscape in Guanzhong area.

At present, there are few researches on the application of text mining technology in speaker relationship recognition and comment sentiment analysis of academic speech. Therefore, this paper refers to the ideas and methods in the field of text mining, and applies it to academic speech to explore and analyze the speaker relationship and comment emotion of academic speech. The specific research ideas of this study is to design and implement the text mining system through python programming that grabs the YiXi website page of the speaker and comment information to store in database. Using relationship network diagram construction, emotional orientation analysis and emotional classification, we attempt to study the speaker characteristics of speech from the perspective of speakers and audiences. It is expected to provide reference for online speech platform to better understand the mechanism behind so as to improve the online speech development environment in smart cities.

3. Data source

In this paper, the data of YiXi Speech on YiXi official website is collected by Web Crawler. By December 12, 2019, the information of 690 speakers and their complete speech texts as well as 23,685 comments have been obtained. Based on the full text information of speeches and related information, a speech information table, a speaker information table and a comment information table have been constructed. The relationship between the three tables is shown in Figure 1. Those tables are connected by the primary key "id". The "speaker" entity publishes a "speech" while they are in a one-to-one relationship. The "speech" entity contains "comments" entity and they have a one-to-many relationship.

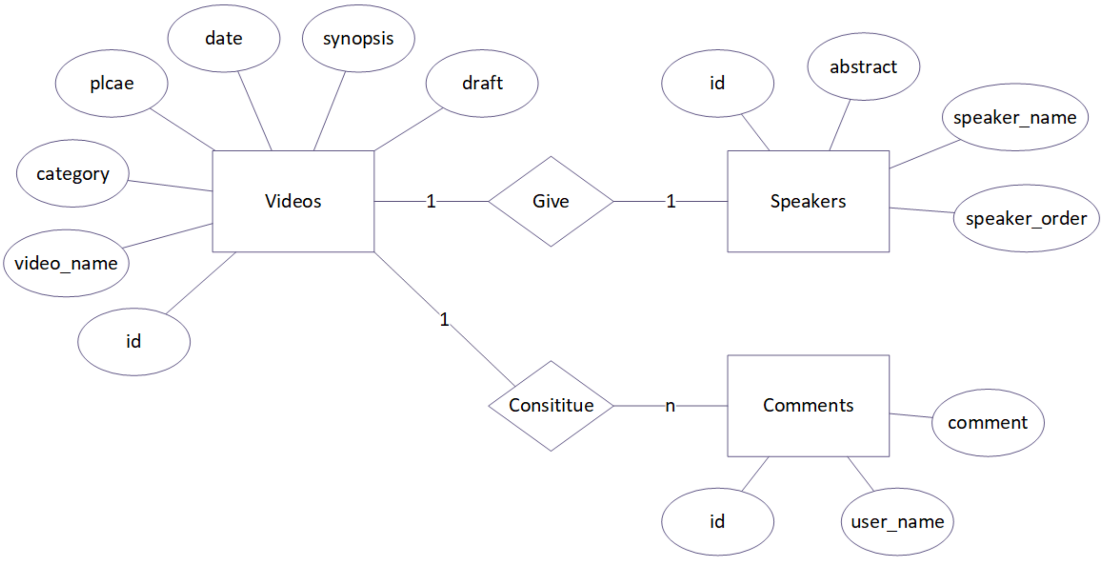

Figure 1. Database E-R diagram.

4. Analysis of speaker relationship network

YiXi invites people who have made outstanding contributions and achievements in their own research fields or have different ideas and concepts as speakers. The speakers are usually good at using their unique perspectives to lead the audience to discover different things. Therefore, the speakers have a certain guidance to the theme of the speech and the evaluation of the audience, and the two are closely related.

In order to discover the relationship between the professional characteristics of the speaker and the topic of the speech, we developed the method based on the co-occurrence analysis. With the help of "graph" theory, we construct and draw the relationship network between the speaker's profession and his topic to analyze the composition characteristics of YiXi speakers, that is, whether the content of speakers of the same profession is similar, or can they cover multiple different fields instead of limited to their own work.

4.1. Establishment of YiXi speaker relationship network

4.1.1. Profession labeling of the speakers

The speaker profile field in the speaker information table is labeled manually. In total, the occupations of 685 speakers are labeled while 118 occupations are marked. Part of the label information is shown in Table 1.

Table 1. Examples of profession labeling information for YiXi speakers.

NO.	Speaker name	Speech category	Speaker profile	First occupation	Second occupation
1	Chen Tongkui	Entrepreneurship	Initiator of the National Returning College Students Forum, Deputy Director of the Social Enterprise Research Center of Shanghai University of Finance and Economics	sponsor	director
2	Liang Hong	Society	Writer, Professor of Chinese Department of China Youth University for Political Sciences	writer	professor
3	Zhu Zhiwei	Design	Chief Type Designer of Founder Font Library	Designer	
4	Zhang Lei	Movie	Director of the movie "August"	director	
5	Liu Shurun	Environment	Botanist, grassland ecologist	Botanist	ecologist
6	Qin Bo	Recording	Reporter, director of the documentary film "Human World"	reporter	director
7	Wang Zhenshan	Entrepreneurship	Microduino's founder and CEO	Founder	CEO
8	Yang Xin	Art	Young artist	artist	
9	Huang Jixin	Entrepreneurship	Co-founder of Know, COO (Chief Operating Officer)	Founder	COO
10	Anja aronowsky cronberg	Culture	Editor-in-chief of independent fashion magazine Vestoj	Editor in chief	

4.1.2. Construction of the co-occurrence matrix of speakers

The identity matrix is created with the speaker's name as the index and list as shown in Eq (1), where S_i represents the name of ith speaker, r_{ij} represents the co-occurrence coefficient of ith speaker and jth speaker.

$$matrix_{co} = \begin{array}{c} \\ S_1 \\ S_2 \\ \vdots \\ S_{685} \end{array} \begin{array}{c} \begin{array}{cccccc} S_1 & S_2 & \cdots & S_j & \cdots & S_{685} \end{array} \\ \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1j} & \cdots & r_{1685} \\ r_{21} & r_{22} & \cdots & r_{ij} & \cdots & r_{2685} \\ \vdots & \vdots & & \vdots & & \\ r_{6851} & r_{6852} & \cdots & r_{685j} & \cdots & r_{685685} \end{pmatrix} \end{array} \tag{1}$$

The co-occurrence logic of occupation and speech category between any two speakers is shown in Table 2. It can be seen that the same category of the speeches given by two speakers will lead to adding one to co-occurrence coefficient of them, while the same occupation of two speakers will result in adding two to co-occurrence coefficient of them.

Table 2. Co-occurrence matrix construct logic.

NO.	Co- occurrence field	Same or not	r_{ij} value in matrix
1	Speech category	Yes	+ 1
2		no	+ 0
3	Speaker occupation	Yes	+ 2
4		no	+ 0

Following the rules specified in Table 2, the element values of the constructed matrix and their specific meanings are shown in Eq (2) where SC_i, SC_j represents the speech category of speaker *i* and speaker *j* respectively, Pro_i, Pro_j represents the profession of speaker *i* and speaker *j* respectively.

$$r_{ij} = \begin{cases} 0 & \text{if } SC_i \,!= SC_j \text{ AND } Pro_i \,!= Pro_j \\ 1 & \text{if } SC_i = SC_j \text{ AND } Pro_i \,!= Pro_j \\ 2 & \text{if } SC_i \,!= SC_j \text{ AND } Pro_i = Pro_j \\ 3 & \text{if } SC_i = SC_j \text{ AND } Pro_i = Pro_j \end{cases} \tag{2}$$

Algorithm 1 gives the method to establish the speaker relationship network step by step. It can be seen that each value in co-occurrence matrix is calculated first of all by comparing speech categories and occupations between each pair of speakers. Secondly, those values and the corresponding speaker names are used to construct edges of the relationship network. Finally, the nodes in the network are built based on the speaker information, including their names, speech categories, etc.

4.2. Speaker relationship network analysis

With the help of Gephi software, the network of speakers is drawn via importing data sets produced in the previous section to the software. By using FR (Fruchterman Reingold [29]) network layout algorithm, the undirected network drawn is shown in Figure 2 where different colors stand for various speech categories.

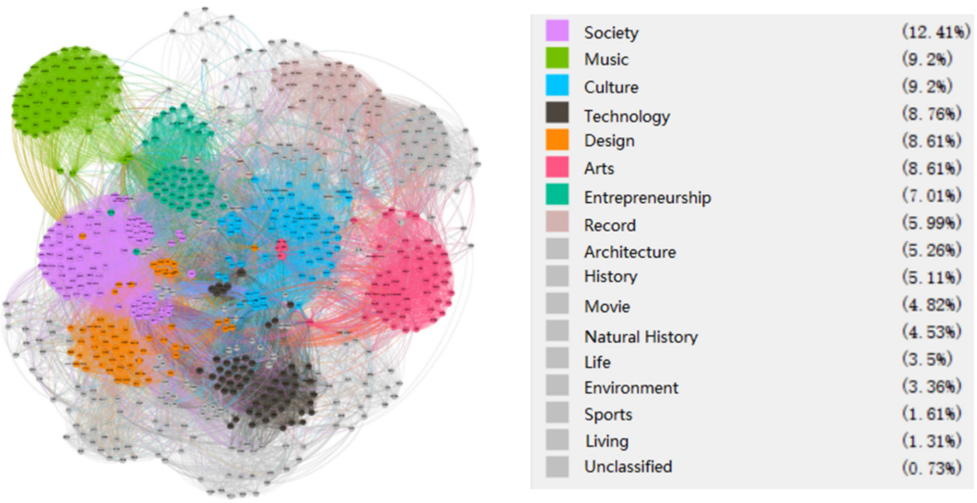

Figure 2. Network diagram of YiXi speakers.

Algorithm 1. Speaker relationship network construction.

Input: professional labeling information set for speakers of YiXi *SpkInfo*

Output: the nodes set *Nodes*, links and associated weights set *Edges* of the speaker relationship network

Procedure:

```
read SpkInfo, each of which is represented as
    $spinfo_i = \{SpkName_i, SphCate_i, SpkProf_i, FirstOccu_i, SecondOccu_i, ThirdOccu_i\}$
    Nodes = {}, Edges = {}
create a matrix whose size is (length of (SpkInfo))* (length of (SpkInfo) )in which all values equal to zero
for i in [1, length of (SpkInfo)]
    $j = i + 1$
    for j in [j, length of (SpkInfo)]
        if $SphCate_i == SphCate_j$, matrix[i][j] = matrix[i][j] +1, matrix[j][i] = matrix[i][j]
        endif
        if $FirstOccu_i$ in $\{FirstOccu_j, SecondOccu_j, ThirdOccu_j\}$ or
            $SecondOccu_i$ in $\{FirstOccu_j, SecondOccu_j, ThirdOccu_j\}$ or
            $ThirdOccu_i$ in $\{FirstOccu_j, SecondOccu_j, ThirdOccu_j\}$
        matrix[i][j] = matrix[i][j] +2, matrix[j][i] = matrix[i][j]
        endif
    endfor
endfor
extract lower triangle of matrix as matrixlt
for i in length of matrixlt.rows
    $j = i + 1$
    for j in length of matrixlt.columns
        if matrixlt[i][j] != 0
        create $edge_i = \{SpkName_i, SpkName_j, 'undirected', matrixlt[i][j], remark_i\}$
        add $edge_i$ to Edges
        endif
    endfor
endfor
extract and save first column and second column of Edges to Nodes
search corresponding $SphCate_i$ in SpkInfo with $SpkName_i$ in Nodes
add $SphCate_i$ to Nodes
return Nodes, Edges
```

The network contains 685 nodes and 25,267 edges, which has the characteristics of a complex network. Each node represents a speaker with his speech category as the label and each edge stands for the co-occurrence relationship between two speakers with the times of co-occurrence as the weight of that edge.

The graph characteristics of the YiXi speaker relationship network are computed and listed in Table 3.

It can be seen from Table 3 that the speaker relationship network of YiXi has a certain degree of modularity (0.623). The number of edges connected to each node is large (average degree = 73.772), but the degree of connection between nodes is low (graph density = 0.108). It can be concluded that the classification of speech categories in YiXi is relatively appropriate, and it has a better clustering effect for speakers of the same speech category (average clustering coefficient = 0.772). The difference between different types of clusters is relatively high. Although there are more co-occurring objects for a single speaker, the degree of co-occurrence is low.

Furthermore, the network diagram is redrawn by selecting edges of different weights. As shown in Figure 3, edges with weights 1, 2 and 3 are selected in the relationship network respectively. In Figure 3(a), the network only retains edges with weight 1, which means that the speech category is the same and profession is different where 14,095 (55.78%) edges are visible. In Figure 3(b), the network keeps only the edges with weight 2, that is, the network of speakers with different speech categories and the same profession where 7803 (30.88%) edges are visible. Finally, in Figure 3(c) the network keeps the edges with weight 3, that is, the network of speakers with the same speech category and the same profession where 3369 (13.33%) edges are visible.

Table 3. Characteristics of the network diagram of YiXi Speakers.

Characteristic	Metrics	Meaning	Index value
The internet	Average degree	The average number of edges connected to each node. The higher the degree of a node, the more points are connected to it, and the more critical the node is.	73.772
	Network diameter [30]	The maximum value of the distance between any two nodes.	4
	Graph density	Measurement of the integrity of the network. The larger the result, the tighter the node connection in the graph. A complete graph has all possible connected edges, that is, any two nodes are edge connected, and its density is 1.	0.108
	Modularity [31,32]	Measurement of the quality of community division. High modularity means that the internal link density is high and the external is sparse. Generally, when modularity >0.44 means that the network diagram has reached a certain degree of modularity.	0.623
	Connected component [33]	A maximal connected subgraph in an undirected graph. In order to determine whether one vertex in the graph can reach another vertex, that is, whether there is a path reachable between any two vertices in the graph.	1
node	Average clustering coefficient [34]	The average clustering coefficient gives an overall indication of a node cluster or clique, indicating the degree of interconnection between each node and its surrounding nodes.	0.772
side	Average path length	The average value of the distance between any two nodes which reflects the degree of separation between nodes in the network. The smaller the value, the greater the connectivity of the nodes in the network.	2.252

From the relationship network of the speakers, it can be seen that under the fixed classification of speech categories, there are big differences in the occupations of YiXi speakers after selecting different weights respectively. Meanwhile, it is found that more than half of the speakers come from different professional fields and deliver the same category of speeches. Nearly one-third of the speakers have the same profession but are able to deliver different types of speeches, indicating that selection of YiXi platform in the invitation of speakers is diversified, and the speakers are rich in practice in their professional research and professional experience.

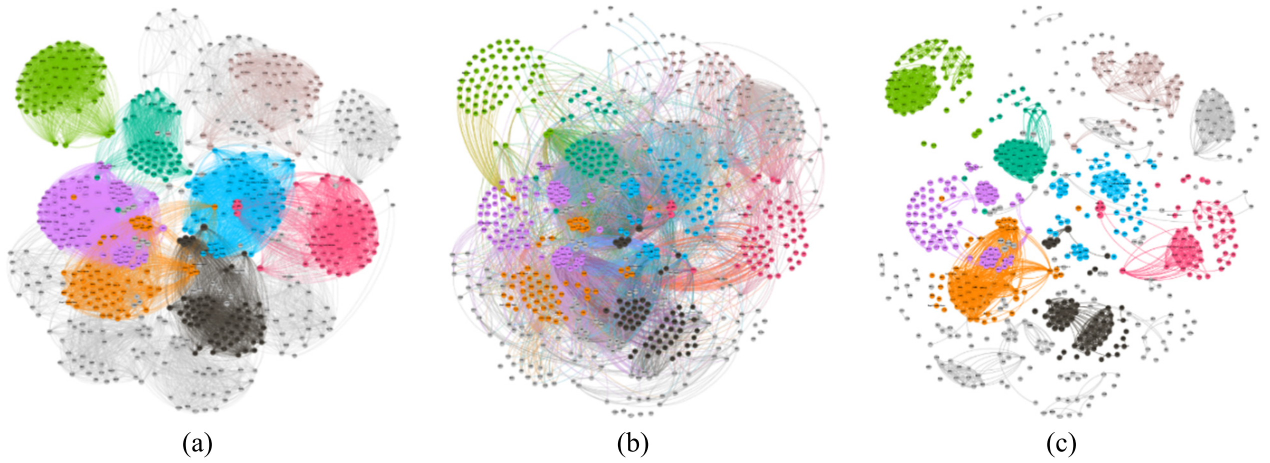

Figure 3. Network graph of speaker relationship (a) edge weight = 1; (b) edge weight = 2; (c) edge weight = 3.

5. Sentiment analysis of speech comments

Emotional analysis is a process of processing and analyzing subjective texts with emotional tendency. There are a large number of user-participated comments under each speech video of YiXi on the content of the speech and the speaker's opinions. How to effectively apply emotional analysis to unstructured data, through emotional rating and emotional segmentation, quantitative description of qualitative data, in order to help the platform better understand the user's point of view, mining the views of the audience, is the main purpose of this study to carry out emotional analysis for speech review. This paper studies the rule-based method and implements the calculation and classification of emotion score by constructing emotion dictionary. This method has the advantages of high precision and strong operability.

5.1. Establishment of dictionaries for emotional analysis

5.1.1. Establishment of emotion vocabulary dictionary

In this paper, "Chinese Emotion Vocabulary Ontology" [35] (hereinafter referred to as vocabulary ontology) is used as the dictionary of emotion words, which was sorted and annotated by the Information Retrieval Research Office of Dalian University of technology. The vocabulary ontology describes Chinese words or phrases from the aspects of property of words, emotional intensity, emotional category and emotional polarity, etc. The example subset of the lexical ontology is shown in Table 4.

The property of words includes noun, verb, adjective, adverb, network word, idiom and preposition. Based on the Chinese traditional emotional expression, the emotional category comprises 7 main categories (that includes Sad, Like, Dislike, Angry, Fear, Surprise, Happy) and 21 subcategories, e.g. under Happy category, there are two subcategories of Joyful (PA) and Relieved (PE). The emotional intensity can be divided into five grades: 1, 3, 5, 7 and 9 where 9 represents the max intensity of feeling and 1 means the minimum. Emotional polarity can be divided into four types: 0, 1, 2 and 3 where 0 is neutral, 1 is commendatory, 2 is derogatory, and 3 can be commendatory or derogatory. In order to facilitate the calculation of emotional score, we change the emotional polarity value of derogatory words to −1, and multiply the emotional intensity to emotional polarity of each word to obtain its emotional value.

Table 4. Example subset of Chinese emotion vocabulary ontology.

Word	脏乱 (messy)	战祸 (war)	易如反掌 (As easy as pie)	英姿飒爽 (bright & brave)
Word Property	adj	noun	adj	idiom
Number of Word meaning	1	1	1	1
Index of Word Meaning	1	1	1	1
Sentiment Classification	NN	ND	PH	PH
Intensity	7	5	9	9
Polarity	2	2	1	1
Auxiliary Sentiment classification		NC	PG	PB
Intensity		5	3	3
Polarity		2	1	1

5.1.2. Establishment of degree adverb dictionary

The degree adverb dictionary used in this study is the “Level Words (Chinese)” [36] in the CNKI sentiment analysis vocabulary set, which contains a total of 219 adverbs that are divided into 6 levels. This paper has customized the emotional intensity of words at different levels, as shown in Table 5.

Table 5. Settings of dictionary of degree adverbs.

No.	Degree level	Emotional intensity	Number of words
1	extreme	1.75	69
2	very	1.5	42
3	more	1.25	37
4	ish	0.75	29
5	insufficiently	0.5	12
6	over	0.25	30

5.1.3. Establishment of stop word dictionary

After researching and comparing different stop vocabularies [37], this article merges the texts in the three lists of Baidu stop vocabulary, Harbin Institute of Technology stop vocabulary and Sichuan University’s machine Intelligence Laboratory stop vocabulary. To avoid the situation where some emotion words, negative words and degree adverbs are filtered out when removing stop words, a total of 2837 stop words that do not contain those three types of words are selected through text deduplication.

5.1.4. Establishment of negative word dictionary

This article uses the Chinese negative vocabulary provided by the CSDN (Chinese Software Developer Network) community, which contains 71 negative words. Since some negative words overlap with emotion words and degree adverbs, which affects the calculation of the final emotion score, 60 negative words are finally selected through text de-duplication.

5.2. Establishment of dictionaries for emotional analysis

The process of sentiment analysis is shown in Figure 4. Algorithm 2 gives the overall process of the sentiment analysis that consists of Algorithms 3–5. At the beginning, it receives original comment text set from database in which the web crawler stores data after collecting information from YiXi official website. Then, Algorithm 3 provides a method to transform comment texts to separate sorted words so as to facilitate the following sentiment calculation. In Algorithm 4, a calculation method is proposed to obtain sentiment score of each comment and each speech by applying corresponding dictionaries to processed words provided by Algorithm 3. Finally, using sentiment scores calculated in Algorithm 4 and 5 classifies the comments into 7 main emotional tendency categories defined in emotional dictionary. The details of algorithms are discussed in the following sections.

Algorithm 2. Overall sentiment analysis process.

```
Input: original comment text set $C$
Output: processed comment text set $CP$ including sentiment analysis results
Procedure:
1. establish stop word dictionary, negative word dictionary, degree adverb dictionary, emotion word dictionary and
   emotional category dictionary
2. extract original comment text $C$ from database, $C = \{c_1, c_2 \ldots c_i, \ldots c_{23685}\}$, set $CP = \{\}$
3. for all $c_i \in C$
4.     run Algorithm 3 to process $c_i$ to word set $w_i = \{id, w_{i1}, w_{i2} \ldots w_{ij}, \ldots w_{in}\}$ where $n$
        represents
        the number of sentences in a comment
5.     run Algorithm 4 to calculate sentiment tendency score $se_i$ of $c_i$, to calculate sentiment tendency score $sp_i$ of
        a speech which consists of a group of $c_i$ that has the same speech (group) ID.
6.     add $se_i$ to $c_i$
7.     if $c_i$ is the last comment in the group, add $sp_i$ to $c_i$
8.     endif
9.     run Algorithm 5 to calculate sentiment category $ca_i$ of $c_i$
10.    add $ca_i$ to $c_i$
11.    add $c_i$ to $CP$
12. endfor
13. return $CP$
```

Algorithm 3 implements the data preprocess function that converts original comment text to sorted words. There are three hierarchical elements during the preprocess algorithm as shown in Eq (3). It can be seen that the comment set is segmented to comments, sentences and words.

After the process of segmentation, stop words in the dictionary are retrieved from the word dataset. If there is a match, the corresponding stop words are deleted from the word dataset. Finally, the output of algorithm 3 is the processed word set each of which is connected the original comment by comment id.

Algorithm 4 is the core algorithm in the process whose task is to receive the previous word set and calculate the sentiment score of comments and speeches. It is assumed that the sentiment score of a comment is the average score of the sentences that it contains. Moreover, the sentiment score of a sentence is the weighted sum of all words in it. Hence, the mathematical model can be established in Eq (4), where *se* stands for the sentiment score of a comment, *n* represents the number of emotion words in each sentence of a comment, *m* represents the number of sentences in a comment.

$$\text{comment set } C=\begin{cases}\text{comment } c_1=\begin{cases}ss_{11}\\ ss_{12}=\begin{cases}w_{121}\\ w_{122}\\ \vdots\end{cases}\\ \vdots\end{cases}\\ \vdots\\ \text{comment } c_i=\begin{cases}\text{sentence } ss_{i1}\\ \vdots\\ \text{sentence } ss_{ij}=\begin{cases}\text{word } w_{ij1}\\ \vdots\\ \text{word } w_{ijk}\\ \vdots\end{cases}\\ \vdots\end{cases}\\ \vdots\\ \text{comment } c_{23685}=\begin{cases}ss_{236851}\\ ss_{236852}=\begin{cases}w_{2368521}\\ w_{2368522}\\ \vdots\end{cases}\\ \vdots\end{cases}\end{cases} \tag{3}$$

$$se=\frac{\sum_{j=1}^{m}\sum_{i=1}^{n} weight_i * intesity_i * polarity_i}{m} \tag{4}$$

Algorithm 3. Convert comment text to word set.

Input: comment $c_i \in C$

Output: word set $w_i = \{id, w_{i1}, w_{i2} \ldots w_{ij}, \ldots w_{in}\}$

Procedure:

```
read $c_i \in C$ where $c_i=\{id, comment\}$, set $w_i = \{\}$
read Stop word dictionary $SD=\{sw_1, sw_2, \cdots sw_i, \cdots sw_{2837}\}$
extract comment from $c_i$
split comment to sentence set $SS_i=\{ss_{i1}, ss_{i2}, \cdots ss_{ij}, \cdots ss_{in}\}$ by punctuations
for all $ss_{ij} \in SS_i$
    split $ss_{ij}$ to word subset $w_{ij} = \{w_{ij1}, w_{ij2} \ldots w_{ijk}, \ldots w_{ijn}\}$ by using word segment
      method provided by "jieba" Python extended package
    for all $w_{ijk} \in w_{ij}$
        delete $w_{ijk}$ if $w_{ijk}$ in SD
    endfor
    add $w_{ij}$ to $w_i$
endfor
add id from $c_i$ to $w_i$
return $w_i$
```

Algorithm 4 Calculate sentiment score of each comment and each speech

Input: word set $w_i = \{w_{i1}, w_{i2} \ldots w_{ij}, \ldots w_{in}\}$

Output: se_i: sentiment score of c_i, sp_i: sentiment score of a speech

Procedure:

```
Read Negative word dictionary ND, Degree adverb dictionary DD and Emotion word dictionary ED
set $se_i = 0$; $sp_i = 0$; $se\_set = \{\}$
for all $w_{ij} \in w_i$
    for all $w_{ijk} \in w_{ij}$
        if $w_{ijk}$ in ED, acquire the intensity and polarity of the similar word,
        $value_{ijk} = intesity * polarity$
        endif
        if $w_{ijk}$ in ND, $value_{ijk} = -1$
        endif
        if $w_{ijk}$ in DD, $value_{ijk} = degree$
        endif
        create dictionaries to store the index, word values and types
      $value = \{\{'ij1', value_{ij1}, type_{ij1}\}, \ldots \{'ijk', value_{ijk}, type_{ijk}\} \ldots \{'ijn', value_{ijn}, type_{ijn}\}\}$
         where n represents the number of the words in a sentence
    endfor
endfor
for all $w_{ij} \in w_i$
    set variable $coeff = 1, sentence\_value = 0$
    for all $w_{ijk} \in w_{ij}$
        if $type_{ijk}$=ND or DD, $coeff = coeff * value_{ijk}$
        endif
        if $type_{ijk}$=ED, $sentence\_value = sentence\_value + coeff * value_{ijk}$
        set variable $coeff = 1$
        endif
    endfor
    add $sentence\_value$ to se_set
endfor
$se_i = average(se\_set)$
$sp_i = average(se_i)$ where $se_i$ belongs to the same speech (group)
endfor
Return $se_i$, $sp_i$
```

In addition, $weight_i$ is decided by the negative words and degree adverbs according to the corresponding dictionaries. $intesity_i$ and $polarity_i$ of an emotion word can be found in emotion word dictionary. Algorithm 4 finally gives the sentiment score of each comment and takes the comments that belongs to the same speech to obtain the sentiment score of each speech.

In order to acquire the emotional tendency of each comment, Algorithm 5 first of all establishes emotional category dictionary based on emotion word dictionary. Then it searches for the most similar category based on the difference calculated based on Eq (5), where *seca* represents the sentiment score under each category defined in emotional category dictionary. The calculation process of *seca* follows the procedure of Algorithm 4.

$$Difference = |se - seca| \tag{5}$$

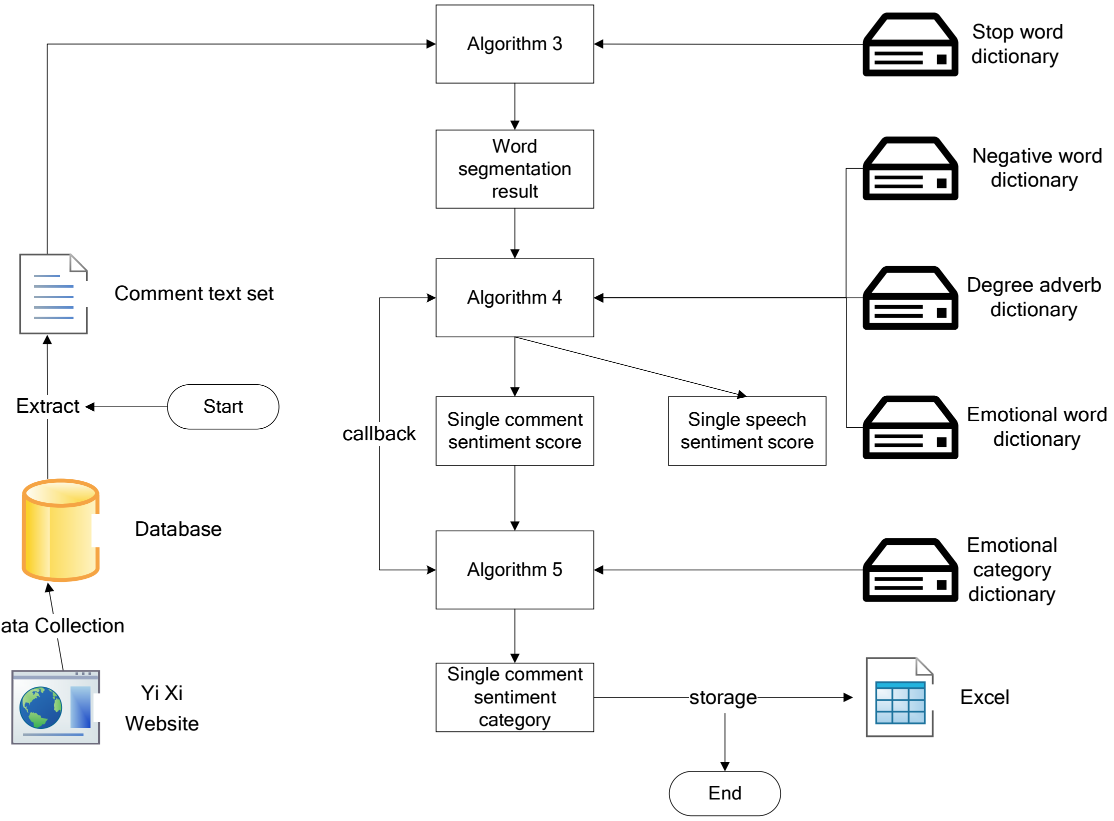

Figure 4. The workflow diagram of Algorithm 2.

Algorithm 5. Calculate comment sentiment category.

Input: se_i: sentiment score of c_i, word set $w_i = \{w_{i1}, w_{i2} \ldots w_{ij}, \ldots w_{in}\}$

Output: ca_i: sentiment category of c_i

Procedure:

```
1. read Emotional category dictionary $ECD$
2. for all categories $cate_i$ in $ECD$
3.     callback sentiment score calculation algorithm in Algorithm 4 by replacing $ED$ with $cate_i$
       to get sentiment score $ca\_se_i$ of $cate_i$
4. endfor
5. create $cate\_score_i = \{'cate_1':ca\_se_1, \ldots 'cate_i':ca\_se_i, \ldots 'cate_7':ca\_se_7\}$
6.   for all $'cate_i':score_i$ key-value pairs in $cate\_score_i$
7.       obtain the $'cate_i':score_i$ pair that produces $\min(abs(se_i - score_i))$
8.   endfor
9.   $ca_i = cate_i$
10. return $ca_i$
```

5.3. Sentiment results analysis

Descriptive statistics and regression analysis are performed on 23,685 comment texts under 690 speech videos, and graphs are drawn. The sentiment classification histogram of all comment texts is shown in Figure 5.

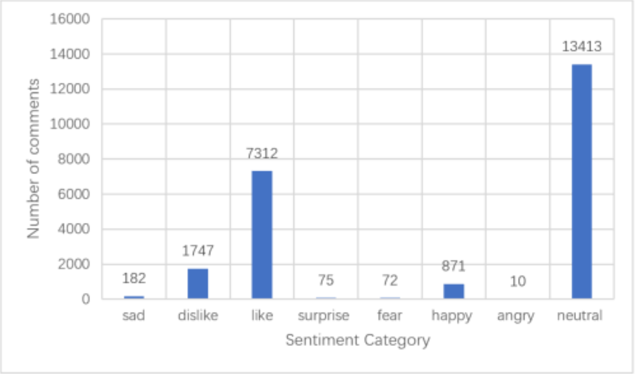

Figure 5. Overall sentiment classification of YiXi comments.

It can be seen that excluding the "neutral" category, the number of comments with the emotional category "like" is the most, with 7312. The emotional category with "angry" has the least number of comments, with only 10. The number of comments on "dislike", "happy" and "sad" are 1747, 871 and 182 respectively. The number of comments in the rest of the sentiment categories are below 100 and the difference between them was small. Furthermore, the positive emotional tendency that includes "like" and "happy" is approximately over 80% without counting "neutral" category. Therefore, besides the popularity, the effect of YiXi speeches is affirmed.

The scatter charts of emotional score of all comments and speeches are shown in Figures 6 and 7 respectively.

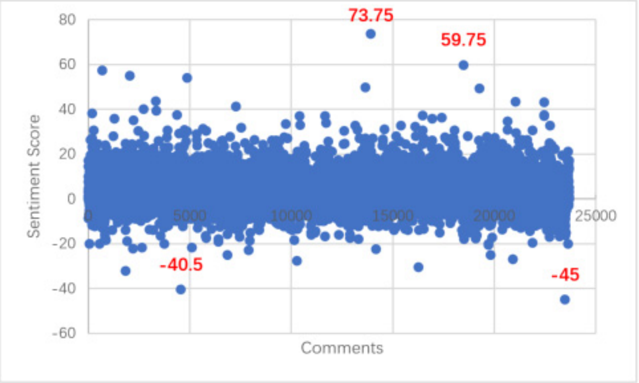

Figure 6. Scatter plot of emotional scores of YiXi comments.

From Figure 6, it can be seen that the general trend of the sentiment tendency of comments is a horizontal line where positive comments and negative comments are symmetrically distributed along the trend line. Around 53% of the comments have a zero sentiment score value which means that half of the comments show neutral attributes. Generally, the average value of emotional scores of comments is about 1.94 which means all comments demonstrate a slightly positive emotional tendency. As shown in Figure 7, it can be seen that around 86% (594 speeches in 688) of the speeches have positive sentiment score. The statistics of both figures are listed in Table 6. Since the sentiment score of a speech is the average value of the comments under that speech, the mean values of both figures are very close. Unlike the distribution of comments' sentiment scores, the scores of speeches reside in a smaller range that has a much less standard deviation.

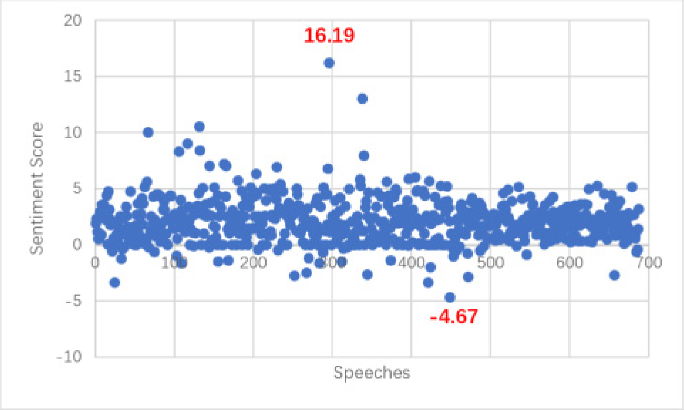

Figure 7. Scatter plot of emotional scores of YiXi speeches.

Table 6. Statistics of sentiment scores.

	Maximum	Minimum	Mean	Standard deviation
Sentiment score of comments	73.75	−45	1.94	5.15
Sentiment score of Speeches	16.19	−4.67	2.06	1.94

Analysis of YiXi comments and speeches. The content of the 4 comments marked by the data tags in Figure 6 is shown in Table 7. The speeches with the highest and lowest sentiment scores are listed in Table 8 and the corresponding comment content is shown in Table 9. All comments in Tables 7 and 9 are translated from Chinese language.

From the figures and tables below, it can be found that unlike e-commerce product reviews, film and television reviews, etc., the content of YiXi speech reviews is more about the audience's feelings after "reading" the speech in the form of video, text or audio. The original intention of the review is "reading feeling" or "personal perception", instead of a direct evaluation of speech quality.

From the perspective of the emotional tendency and emotional category of the comments, the users of YiXi generally hold a positive attitude towards the experiences and views described in YiXi Speech, and the user reviews are mixed. There is no affirmation or negation of blindly following the crowd. As to the specific content of the comments, users can objectively, rationally and truthfully express their attitudes towards the views and things shared by the speaker, and further express their personal views clearly from their own perspective. The comments contain rich content and implication, and use appropriate words, showing the unique thinking mode of the audience.

Table 7. Extreme values in the overall sentiment tendency of YiXi Comments.

Speech id	Topic	Comments	Sentiment score	Emotion category
S_706_20180916	The disappearing classical gardens	What is important is the inheritance and development of Chinese culture and some innovation, including various gardens and other Chinese arts, which are the embodiment of Chinese national thought and the carrier of Chinese culture. People like Mr. Chen and other teachers engaged in the restoration of cultural relics and culture in this area are worthy of our respect. The ideal state of Chinese culture may be that it retains the integration and complementation of those truly excellent and meaningful Chinese culture that affects the national character, the new excellent culture with Chinese characteristics and the culture with Chinese characteristics in the future, and then develop continuously to form our own unique cultural style of the Chinese nation. No matter whether it is exactly the same, as long as it is really conducive to the inheritance of national culture, it is worth us to explore and carry forward. The most important thing is to let the real Chinese culture not be cut off.	73.5	like
S_793_20190420	All soulless song of no soul	For rock and roll, maybe I prefer rap. After listening to this speech, which is not a speech, it is really touching, especially when he sings Wahaha, he sings art and life so incisively and vividly, telling his life track with a kind of shouting voice. Similarly, he also tells me who can be called a musician. Compared with stars, they are the products of art, as he said, a group of 18-or-19-year-old people get together to make music. Their hair is gray and their face is more wrinkled. Instead, they are satisfied. They tell their experiences in Tibet in light words, such as “gambling shop” and “three-point style” with machetes. In the pilgrimage to Gangrenboqi, his soul is saved. Wild dogs are accompanied and world of luxurious livings. He is the only one who can sing this out. He pursues the organic combination of business and art in the film, but he is so infatuated with it and struggles for it in music. Even if it is finally some songs without soul, some people will sing and appreciate it.	59.75	like
S_856_20191123	Wild animals or commodities?	The photographer's record is powerful. When the numbness of human beings and the numbness of animals due to cruelty are combined into photographic images, tension will follow. Humans are so smart that they have to explore countless unknowns. Human beings are so conceited that they touch the bottom line of nature. Once the balance is broken, it will bear its own fruit.	−45	dislike
S_338_20170610	How to face love in the golden year of being single	After listening to this issue many times, I still feel very good. But life is too contradictory, when you want to love, the cruel reality of life make people do not dare to step out of that step.	−40.5	sad

Table 8. YiXi speech with the highest and lowest emotional score.

Speech id	Topic	Category	Number of comments	Mean sentiment score	Positive comments	Negative comments
S_350_20151101	firewood	Arts	4	16.19	3	0
S_543_20120826	Has the information revolution ever existed?	Technology	3	−4.67	0	3

Table 9. Comment content of speeches with the highest and lowest emotional score.

Speech	No.	Comments	Sentiment score	Emotion category
S_350_20151101	1	Where can there be a master in Liaoning, I want to learn.	0	neutral
	2	Revealing truth, simplicity, simplicity and the best love.	30.75	like
	3	(⊙o⊙)Wow, like.	5	like
	4	I like this kind of very simple and peaceful artist, a little bit of work and a little bit of perfection.	29	like
S_543_20120826	1	It is society that shapes technology, not technology that shapes society.	−5	dislike
	2	Even if the point of view is correct, negative expression will cause discomfort to the listener. We choose technology, and we are changed by technology.	−4	dislike
	3	Technology has always been used and used by those who need it. The right to choose technology in social progress lies with the decision-makers at that time. For example, the nuclear power technology we use now is not the best, but it is used most.	−5	dislike

6. Conclusions

Based on big data analysis theory and method, this paper uses the text data of YiXi website as the data source and analyzes the relationship network of speakers and sentiment tendency of speech review texts. The main research conclusions may include:

1) There are great differences in profession and speech category among YiXi speakers, which makes the composition of speakers diversified. It reflects the rich practicality of speakers in their professional research and professional experience. Their unique perspectives and model of thinking can lead the audience to reshape the understanding and impression of things, bringing more innovative elements to the content of YiXi speech. It also helps the platform to select the speakers more clearly.

2) The emotional polarity and category of YiXi speech comments are relatively positive and optimistic, and many user comments are objective and true, which has played a good feedback adjustment role for the platform's selection and optimization of speech content.

As an example of application, Ping An International Smart City Technology Company has developed a sophisticated application named ZHINIAO for on-line vocational education [38]. To optimize user experience, it has employed this research in ASKBOB Intelligent Learning Assistant production development. ZHINIAO’s ASKBOB obtains knowledge from various online vocational education course records so as to provide precise searching service and personalized course recommendation services. Using analysis algorithms, ASKBOB could accurately capture the characteristics of online-course audiences and analyze knowledge dissemination among audiences to improve user experience.

The research of this article still has some deficiencies in some aspects. In sentiment analysis, it is difficult to construct an emotional dictionary for texts in different fields. The number of sentiment vocabulary is limited, which leads to partial inaccurate “0” values in the calculation of sentiment scores. In speaker relationship network analysis, speak manual labeling of speaker occupations is time-consuming and labor-intensive, and there are too many labeling categories, which makes the co-occurrence relationship between speakers more complicated and the integrity of the constructed network is not high.

In the follow-up related research, it will be further improved and deepened. The multilingual and multiple category speech information will be expanding, such as analysis based on “TED Talks” and other academic speech data. In the construction of dictionaries, we should continuously enrich and improve the bottom level dictionaries of emotional analysis, and adopt more scientific and standardized professional tagging methods to optimize the annotation results. Meanwhile, it can not only expand the perspective and depth of speech pattern analysis, but also reduce the dependence of program tagging on manual tagging corpus. It may change from supervised learning to unsupervised learning and explore the application of machine learning method in comment sentiment analysis. For example, ZHINIAO also maintains a product named “Intelligent Training-Bot”, which helps users learn the professional skills from corresponding courses. In the future, analyzing of users’ behavior and emotion during the process of learning will be meaningful and challenging.

Acknowledgments

The research is supported by the high level introduction of talent research start-up fund of Nanjing Institute of Technology (Grant NO.: YKJ201989); Open Research Fund of NJIT Institute of Industrial Economy and Innovation Management (Grant NO.: JGKB202001, JGKC 202003); Innovation Fund General Project I of Nanjing institute of technology (Grant NO.: CKJB202003); University Philosophy and Social Science Research Project of Jiangsu province (Grant NO.: 2019SJA2274); Major Project of Philosophy and Social Science Research in Universities of Jiangsu Province Education Department (Grant NO.: 2020SJZDA069). ZHINIAO -ASKBOB Project of Vocational Education Department of Ping An International Smart City Technology Co., Ltd

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. S. A. Rani, M. A. Jabar, R. Abdullah, Y. Y. Jusoh, *The Influence Factors of Knowledge Resilience in Sustaining Knowledge Network in Smart Cities Environment*, 2020 6th International Conference on Information Management (ICIM), London, United Kingdom, 2020.
2. E. Negre, C. Rosenthal-Sabroux, M. Gascó, *A Knowledge-Based Conceptual Vision of the Smart City*, 2015 48th Hawaii International Conference on System Sciences, Kauai, 2015.

3. A. Elabora, M. Alkhatib, S. S. Mathew, M. El Barachi, *Evaluating Citizens' Sentiments in Smart Cities: A Deep Learning Approach*, 2020 5th International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 2020.
4. Y. Cai, Y. Zhao, J. Yang, C. Wang, *A Bus Passenger Flow Estimation Method Based on POI Data and AFC Data Fusion*, ICBDS 2019, Communications in Computer and Information Science, 2019.
5. T. Sun, Y. Zhao, Z. Lian, *People Flow Analysis Based on Anonymous OD Trip Data*, ICBDS 2019, Communications in Computer and Information Science, 2019.
6. Q. Ye, J. Yang, F. Liu, C. Zhao, N. Ye, T. Yin, *L1-Norm Distance Linear Discriminant Analysis Based on an Effective Iterative Algorithm*, IEEE Transactions on Circuits and Systems for Video Technology, 2018.
7. Q. Ye, Z. Li, L. Fu, Z. Zhang, W. Yang, G. Yang, Nonpeaked discriminant analysis for data representation, *IEEE Trans. Neural Networks Learn. Syst.,* **30** (2019), 3818–3832.
8. L. Fu, Z. Li, Q. Ye, H. Yin, Q. Liu, X. Chen, et al., Learning robust discriminant subspace based on joint L2,p- and L2,s-Norm distance metrics, *IEEE Trans. Neural Networks Learn. Syst.*, (2020), forthcoming.
9. J. Lin, *Social Network Analysis: Theory, Method and Applications*, First edition, Beijing Normal University Press, Beijing, 2009.
10. V. Colizza, A. Flammini, M. A. Serrano, A. Vespignani, Detecting rich-club ordering in complex networks, *Nat. Phys.*, **2** (2006), 110–115.
11. S. P. Borgatti, A. Mehra, D. J. Brass, G. Labianca, Network analysis in the social sciences, *Science*, **323** (2009), 892–895.
12. M. Girvan, M. E. J. Newman, Community structure in social and biological networks, *PNAS*, **99** (2002), 7821–7826.
13. M. E. J. Newman, Fast algorithm for detecting community structure in networks, *Phys. Rev. E*, **69** (2004), 066133.
14. V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, *J. Stat. Mech. Theory Exp.*, **2008** (2008), P10008.
15. L. Waltman, N. J. Van Eck, A smart local moving algorithm for large-scale modularity-based community detection, *Eur. Phys. J. B*, **86** (2013), 471.
16. L. Erhan, M. Ndubuaku, E. Ferrara, M. Richardson, D. Sheffield; F. J. Ferguson, et al., Analyzing objective and subjective data in social sciences: implications for smart cities, *IEEE Access*, **7** (2019), 19890–19906.
17. C. Xu, Z. Guan, W. Zhao, Q. Wu, M. Yan, L. Chen, et al., Recommendation by users' multi-modal preferences for smart city applications, *IEEE Trans. Ind. Inf.,* **17** (2021), 4197–4205.
18. N. Kilicay-Ergin, A. S. Barb, *Smart City Document Evaluation to Support Policy Analysis*, 2020 IEEE International Systems Conference (SysCon), Montreal, QC, 2020.
19. K. S. Oza, P. G. Naik, Prediction of online lectures popularity: a text mining approach, *Procedia Comput. Sci.*, **92** (2016), 468–474.
20. A. Ghose, P. G. Ipeirotis, *Designing Novel Review Ranking Systems: Predicting Usefulness and Impact of Reviews*, International Conference on Electronic Commerce. ACM, 2007.
21. S. M. Mudambi, S. David, What makes a helpful online review? a study of customer reviews on Amazon.com, *MIS Q.*, **34** (2010), 185–200.

22. J. Chen, J. Zhang, Y. Zhang, Impact factors of online customer reviews usefulness: a text semantics approach, *Libr. Inf. Serv.*, **56** (2012), 119–123.
23. G. Yin, W. Liu, S. Zhu, What makes a helpful online review? The perspective of information adoption and social network, *Libr. Inf. Ser.*, **56** (2012), 140–147.
24. A. Founoun, A. Hayar, A. Haqiq, *The Textual Data Analysis Approach to Assist the Diagnosis of Smart Cities Initiatives*, 2019 IEEE International Smart Cities Conference (ISC2), Casablanca, Morocco, 2019.
25. H. Yu, Z. Li., Y. Jiang, *Using GitHub Open Sources and Database Methods Designed to Auto-Generate Chinese Tang Dynasty Poetry*, Communications in Computer and Information Science, 2020.
26. M. Zhang, Sentiment analysis of E-commerce reviews based on text mining, *Ind. Sci. Tribune*, **19** (2020), 63–64.
27. X. Li, W. Yu, Text mining of comment dada on video electronic product based on sentiment analysis and relational network, *Intell. Explor.*, **2018** (2018), 1–5.
28. X. Wu, X. Li, W. Zhao, Research on cultural landscape patterns of Guanzhong area based on text mining of Tang poetry, *Landscape Archit.,* **26** (2019), 52–57.
29. T. M. Fruchterman, E. M. Reingold, Graph drawing by force-directed placement, *Software Pract. Exp.*, **21** (1991), 1129–1164.
30. U. Brandes, A faster algorithm for betweenness centrality, *J. Math. Soc.*, **25** (2001), 163–177.
31. V. D Blondel, J. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, *J. Stat. Mech. Theory Exp.,* **2008** (2008), P1000.
32. R. Lambiotte, J. C. Delvenne, M. Barahona, Laplacian dynamics and multiscale modular structure in networks, preprint, arXiv:0812.1770.
33. R. Tarjan, Depth-first search and linear graph algorithms, *SIAM J. Comput.*, **1** (1972), 146–160.
34. M. Latapy, Main-memory triangle computations for very large (Sparse (Power-Law)) graphs, *Theor. Comput. Sci.*, **407** (2008), 458–473.
35. L. Xu, H. Lin, Y. Pan, H. Ren, J. Chen, Constructing the affective lexicon ontology, *J. China Soc. Sci. Tech. Inf.*, **27** (2008), 180–185.
36. Collection of words for sentiment analysis (beta version), 2020. Available from: http://www.keenage.com/html/c_bulletin_2007.htm.
37. G. Qin, S. Deng, H. Wang, Chinese stopwords for text clustering: a comparative study, *Data Anal. Knowl. Discovery*, **1** (2017), 72–80.
38. Ping An Zhiniao Education Platform, 2020. Available from: https://www.zhiniao.com/platform.html#.