A Lightweight Sentiment Analysis Method

YU Qingshuang[1], ZHOU Jie[2], and GONG Wenjuan[1]

(1. China University of Petroleum (East China), Qingdao, Shandong 266000, China;
2. Operation Coordination Department of Tianjin Branch of CNOOC (China) Co., Ltd., Tianjin 300000, China)

Abstract: The emergence of big data leads to an increasing demand for data processing methods. As the most influential media for Chinese domestic movie ratings, Douban contains a huge amount of data and one can understand users' perspectives towards these movies by analyzing these data. In this article, we study movie's critics from the Douban website, perform sentiment analysis on the data obtained by crawling, and visualize the results with a word cloud. We propose a lightweight sentiment analysis method which is free from heavy training and visualize the results in a more conceivable way.

Keywords: web crawler; microblog; text sentiment analysis; word cloud

DOI: 10.12142/ZTECOM.201903002

http://kns.cnki.net/kcms/detail/34.1294.TN.20190920.2106.008.html, published online September 20, 2019

Manuscript received: 2019-05-09

1 Introduction

Text, semantics and social analysis are means to mine users' opinions, the market trend, and other useful information. Currently, text, semantics and social analysis technologies have been widely used in many industries including finance, media and e-commerce. For example, TUDORAN used emotional analysis to show why users block ads [1]; LEE and KWON analyzed the psychology of the people who committed suicide by analyzing the Twitter data, and analyzed the causes so that certain medication is provided to decrease the number of people who committed suicide [2]. Also, in business and governments, people extract useful information that can improve the quality of decisions from massive amounts of data.

Therefore, how to efficiently and accurately analyze the information that reflect people's opinions has become a hot research topic. Recently most researchers use machine learning methods for sentiment analysis. For example, WANG et al. extended the text library by adding network terminology and Wikipedia to the corpus and used it to train convolutional neural networks for text-level sentiment analysis [3]. RASOOL et al. used the convolutional neural network to train the sentiment analysis model to compare the popularity of the two clothing brands on Twitter [4].

However, the cost of sentiment analysis by training the neural network is high: 1) The data set used for training and testing is huge; 2) it takes a long time to train; 3) the development of training models has reached a bottleneck. This is cumbersome for adapting for unseen sentiment analysis and especially not applicable to small business users. HUSSEIN also raised the challenge of sentiment analysis—the accuracy of the analysis still needs to be improved [5]. To this end, we propose a lightweight sentiment analysis method based on SnowNLP that is a Python-written class library for processing Chinese content. SnowNLP performs sentiment analysis by segmenting sentences, part of speech tagging, and emotional judgments.

At the same time, the sentiment analysis methods trained by YANG et al. [3] and RASOOL [4] and others through neural networks cannot visualize the central words for movies and television, which is obviously inconvenient for audiences who want to understand the movie on a glimpse. To this end, we propose to explicitly display the central words of the movie through the combination of Jieba lexicon and word cloud.

We use the proposed method to solve the problem of using neural network to analyze small amount of data while taking time to train on a large dataset. Besides, it displays the central words from the movie review, and opens up new ideas and meth-

ods for the future analysis of small data in sentiment analysis.

2 Basics of Web Crawlers and Sentiment Analysis

2.1 Web Crawlers

The web crawler obtains the source code of the webpage (**Fig. 1**) by using the request library. After obtaining the source code of the webpage, we analyze the source code data. Common data processing methods are used including regular expressions and libraries that extract web page information based on web page node properties, CSS selectors or XPath, such as Beautiful Soup, pyquery, and lxml. These libraries can efficiently and quickly extract useful information from the source code in web pages, such as node attributes, text values, and others. Finally, the processed information is saved for subsequent use.

2.2 Text Sentiment Analysis

With the development of big data, the amount of information on the Internet is increasing and how to easily and accurately extract the information from massive online reviews has become a focus of research in the field of sentiment analysis. However, the text sentiment analysis mostly solves the English text, and is based on social platforms such as Twitter [6]–[8]. For this reason, we propose a Chinese-based sentiment analysis method based on the Chinese text sentiment analysis library.

2.2.1 SonwNLP Library

SnowNLP is a Python-written class library inspired by TextBlob. It can easily handle Chinese text. Unlike TextBlob, Natural language toolkit (NLTK) is not used in SnowNLP, and all algorithms are implemented under the framework. It includes some well-trained dictionaries and can reduce the amount of data set and the amount of training time required for training neural networks.

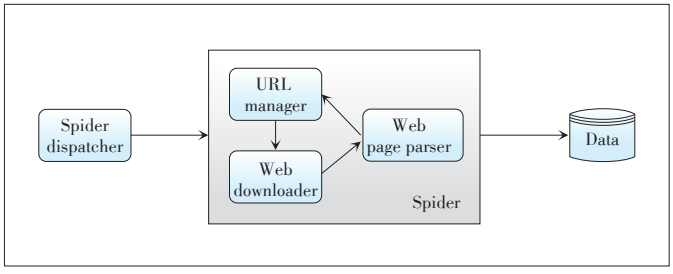
2.2.2 Jieba Word Segmentation

Now, open source tools for Chinese word segmentation include IK (Ik-analyzer), MMseg4j, THU Lexical Analyzer for Chinese (THULAC), Ansj, Jieba, Han Language Processing (HanLP), etc. which are still being updated and maintained. Currently, Ansj, Jieba, and HanLP perform relatively better for segmentation.

Jieba support three word segmentation modes: the precise mode, which attempts to segment the sentence most accurately suitable for text analysis; full mode, which scans all the words in the sentence that can be performed very fast but cannot solve the ambiguity; and search engine mode, which further splits a long word, improves the recall rate, and is suitable for search engine segmentation. It supports traditional word segmentation and dictionaries. The core of the algorithm is to implement efficient word graph scan based on the prefix dictionary and generate a directed acyclic graph (DAG) composed of all possible formation. Dynamic programming is used to find the maximum probability path, and the maximum segmentation combination is found based on word frequencies. For unregistered words, an HMM model based on the formation ability is adopted, and it also uses the Viterbi algorithm.

The Jieba word segmentation method is based on the Trie tree structure to achieve word graph scanning, generates a directed acyclic graph of word formation. Jieba has a dictionary, which contains more than 20 000 words, including the number of occurrences and part of speech. And Trie tree is a well-known prefix tree. If the first few words of a phrase are the same, they have the same prefix, and we can use the Trie tree to store them. It has the advantage of fast search speed.

Data analysis results can be visualized through Jieba word segmentation and a word cloud which makes up for the shortcomings of deep neural networks.



▲**Figure 1. Web crawler work pipeline.**

3 Sentiment Analysis Based on SnowNLP

We will collect the data, process it through the SnowNLP class library, and use SnowNLP to perform word segmentation, part of speech tagging, abstract extraction, and text sentiment analysis. Sentiment analysis simply divides sentences into two categories: positive and negative. A histogram is drawn based on this probability, and the results of the sentiment analysis are explicitly displayed. The obtained data is presented in the central vocabulary of the data set through the Jieba word segmentation and the word cloud.

3.1 Collection of Comments

Comments are collected through the request library. The problem is that some websites will use anti-crawler strategies. The commonly used anti-crawler mechanism is mainly based on the following three types of information: user requests for Headers, user behavior, website directory and data loading methods. We discover that the Douban website uses the anti-crawler based on Headers. It checks and monitors the User-Agent of the Headers. We add Headers directly to the crawler. When crawling, the cookie file after login is saved in the cook-

ie.txt file, and the browser's User-Agent is copied to the Headers. A user-agent is added to the headers and a cookie file is added to break through the anti-crawl mechanism and get the source code.

Firstly we should get the uniform resource locator (URL) of the website and analyze the URL to parse it on a specific page. For the homepage URL of the Douban Movie Critics[1], its prefix[2] is fixed, and the latter 27763742 is the ID of the movie, which is the number given by Douban. We can open the specified page to get the specified ID of the movie by analyzing the page URL of the second page[3].

We discover that the value of the "start" is the number of pages; for example, the second page is 20, the third page is 30, and so on. The value of the "limit" indicates the number of movie reviews displayed on one page. This is the default value and cannot be modified. And the value of the "sort" indicates the latest hotspot comment sorting. And the value of the "status" indicates the comments of users who have seen this movie, but the comments written by users who have not seen this movie are not displayed.

3.2 Data Cleaning

The source code of the crawled webpage is processed, and the useless words are filtered out, leaving only the required information. This process mainly uses the library of bs4 for information processing.

Through the webpage check, we can know that the comments which we want are in the corresponding "p" field in the comment directory. So the latter code parsing is to extract the text contained in this part of code. Since each movie has a lot of reviews, for the timeliness of the information, only the first 200 pieces are obtained for analysis. The iterative processing is performed when the web page is processed, 20 pieces are fetched at a time, and 10 pages can be fetched. We use BeautifulSoup and regular expressions for data cleaning, leaving useful information for subsequent word frequency statistics. Finally, the cleaned string is input into the text file, which is convenient for word frequency analysis and data processing analysis.

3.3 Word Cloud Visualization

The obtained movie reviews are placed in a specific document to facilitate subsequent Jieba word segmentation and displayed in word cloud. After the content is crawled, the next step is to parse and display it. This part of the code applies to any text. For example, it can analyze the movie reviews, as well as the lyrics, news keywords, and so on. The commonly used Jieba function is its cut function, which can divide the entire sentence or the whole paragraph of the article according to the words commonly used in Chinese, so as to facilitate the subsequent analysis of the keywords. In addition, WordCloud is also a great feature that displays the shape of the word cloud in a specific style and displays it in different text sizes depending on how often each word appears. The word cloud outlines include circle, cardioid, diamond, triangle-forward, triangle, pentagon, and star. After the word frequency of each word is obtained, there is a process of sorting the word frequency before the word cloud is displayed. Since the keywords and word frequency are composed of dictionaries, the sorted function is used in the sorting and the keywords are arranged according to the word frequency from large values to small values.

3.4 Part of Speech Analysis

SnowNLP is used to process text. It can be used for word segmentation, annotation, text sentiment analysis, etc. Sentiment analysis classifies text into two categories: the positive and negative. A predicted value of the method is the probability of the emotion.

4 Experimental Results

The movie review information is obtained by the crawler and the acquired information is cleaned by the BeautifulSoup class library. The useful information is retained and stored in the text. The information in the text is processed through the SnowNLP class library to analyze the emotional polarity, and the central subject of the text comment is explicitly displayed through the combination of Jieba word segmentation and word cloud. The experiments are carried out under the following environment: Windows10, Anaconda, WordCloud, Jieba word segmentation, and the SnowNLP class library.

4.1 Crawling Comments

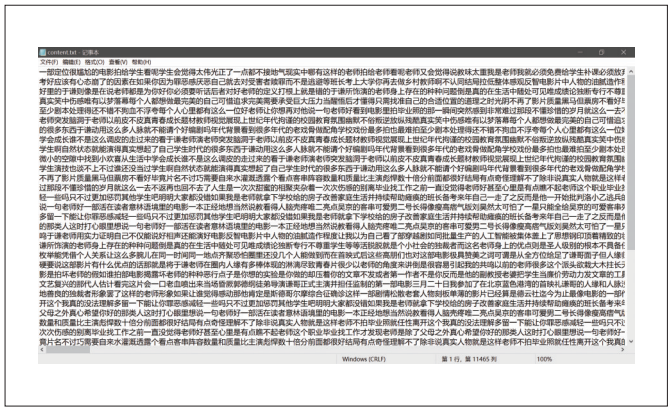
The source code which is cleaned by BeautifulSoup and regular expressions is saved in the contect. txt. An example is shown in **Fig. 2**.

4.2 Data Visualization Analysis

Five categories are picked. They are the animation, suspense, popular, unpopular, and newest. Ten movies are selected from each category. In total, 50 movies are analyzed and the results are visualized in **Tables 1–5.**

From the word cloud of different types of the movies, we found that there are some similarities. The same type of movie word clouds display mostly the same vocabulary, while the different types of movie word clouds mainly show different vocabulary. For example, the most displayed in the word cloud are the types of the movies, such as "animation" and "Cartoon" in the word cloud of the animation type, and "drama" and "suspense" in the word cloud of the suspense type. This is conducive to distinguishing different types of movies, which is of great significance in information recommendation. The word

[1] https://movie.douban.com/subject/27663742/comments?status=P
[2] https://movie.douban.com/subject/
[3] https://movie.douban.com/subject/27663742/comments?start=20&limit=20&sort=new_score&status=P



▲Figure 2. Comments after data cleaning.

clouds can also reflect the audiences' emotional views on this movie, showing the words "like", "exciting", "classic", "junk", and so on. Moreover, there are many names of actors and protagonists in the word cloud of different types of the movies. This could help audiences choose the right movie to watch.

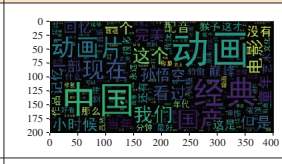
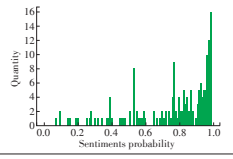
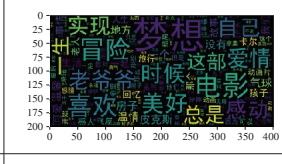
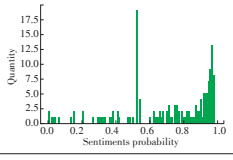
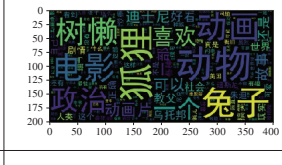
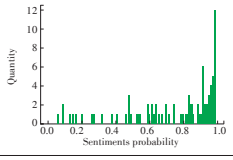
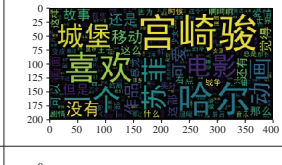
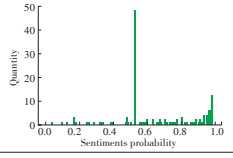
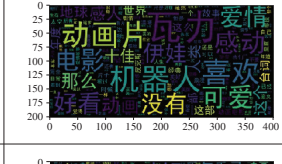
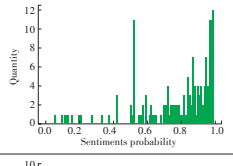
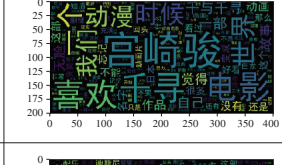
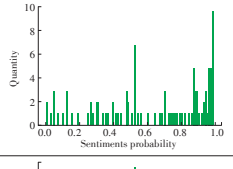
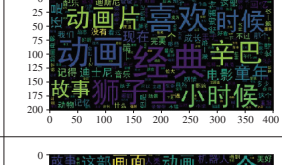
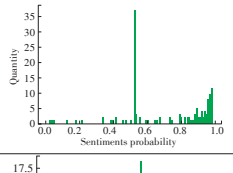
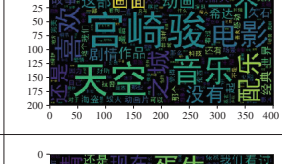
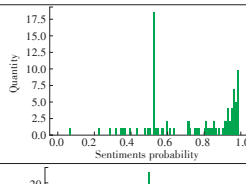
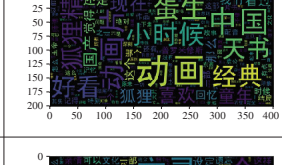
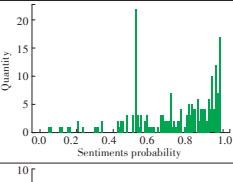
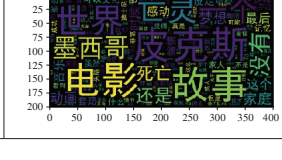
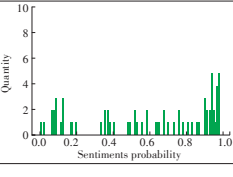
The difference is the emotional analysis of different types of movies. Since our experiment is to randomly select 10 movies on each type, all of them are random, that is, we can treat them as samples. Through the emotional analysis of these samples, we can know the reviews for the animation and popular movies are intensive and mostly positive comments. The suspense movies have many comments but mostly neutral words. The newest movies only have a few and mostly negative comments. Surprisingly, although there are not many movie reviews for the unpopular movies, most comments are positive. By analyzing the sentiment analysis graphs of different types of the movies, cultural-related workers can understand the audiences' preferences for movie types so that they can make corresponding changes.

5 Discussion and Analysis

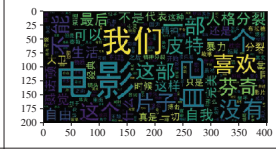
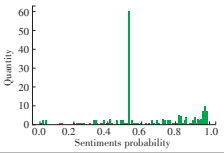
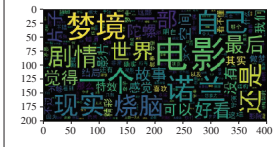
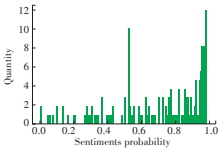
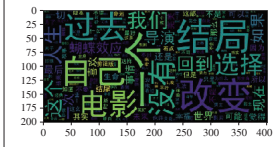
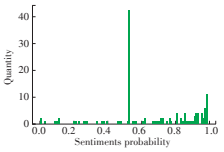
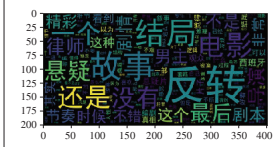
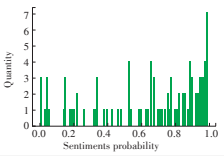
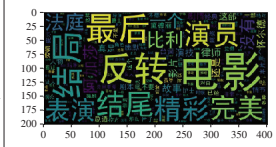
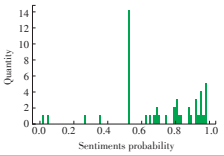
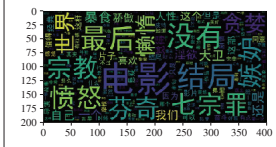
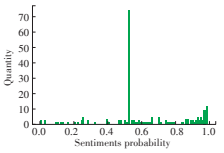
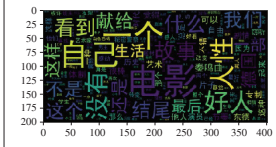
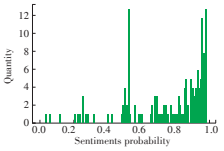
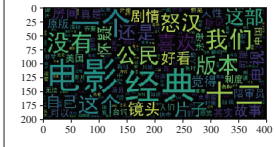
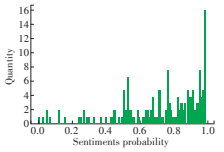
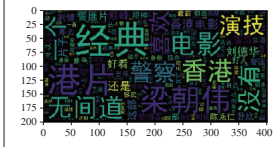
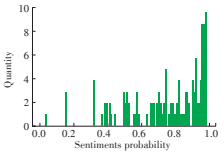
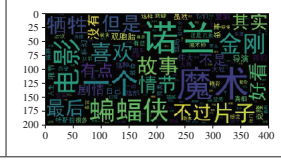
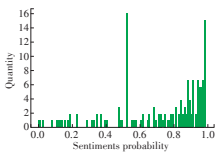
We learned from the experiment results that our method is applicable to different types of movies. It can analyze the expression emotions out of reviews on different types of movies. Combined with the network rating, we can see that the scores of high-score movies could be close to 1 in the sentiment analysis, which indicates that our method is highly credible. And at the same time through the word cloud we can know the central words of reviews on a movie, and it is helpful to conducive to information recommendation. The 50 pictures of the entire program ran down and took only 10 minutes (limited by the speed of the network and the anti-crawling of the website). In general, the proposed method is highly credible, easy to summarize and recommend, and takes less time.

Through this study, the method of sentiment analysis has been explored. From the large amount of data, the long-time machine learning, deep learning and training neural network,

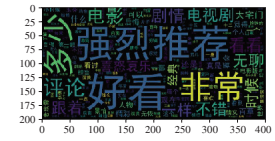
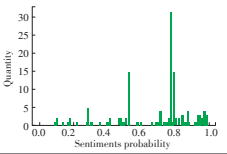
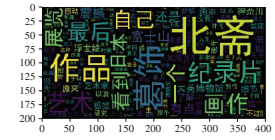
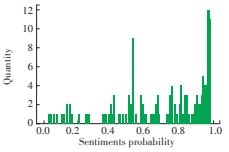
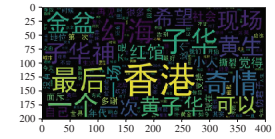
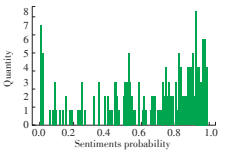
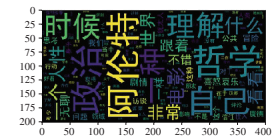
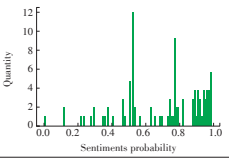
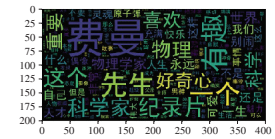
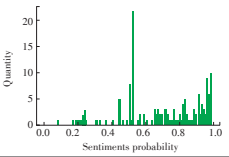
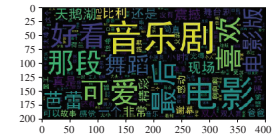
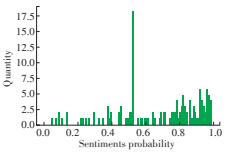
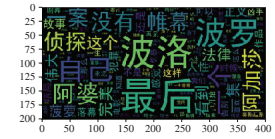
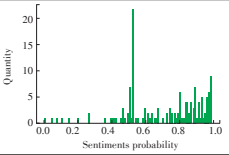
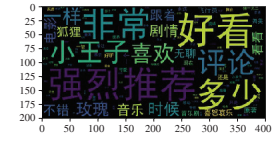
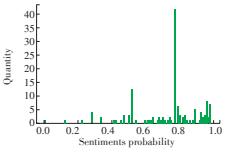
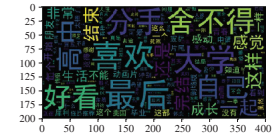
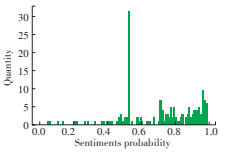
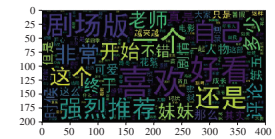
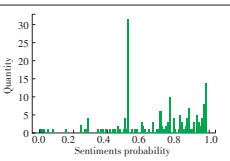
▼Table 1. Visualization analysis of experimental results from the animation movies

Name of the movie	Visualized word cloud	Predicted sentiment probability
The Monkey King		
Up		
Zootopia		
Howl's Moving Castle		
WALL·E		
Spirited Away		
The Lion King		
Laputa: Castle in the Sky		
The Legend of Sealed Book		
Coco		

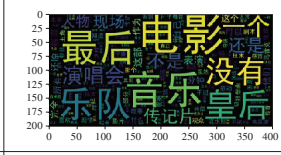
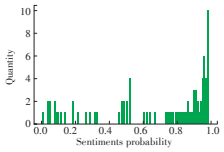
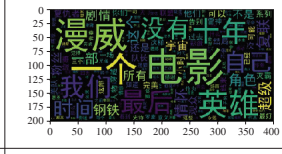
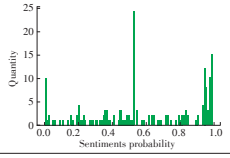
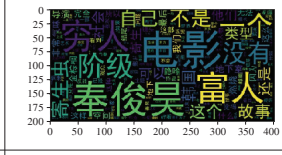
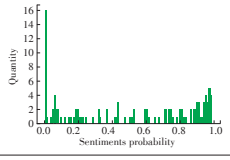
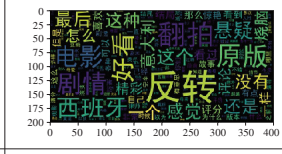
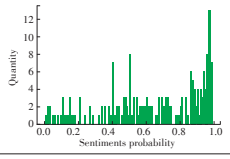
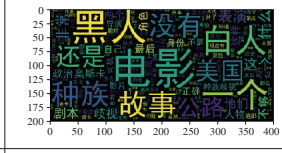
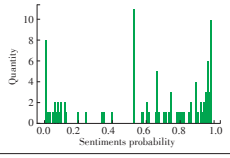
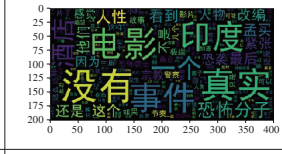
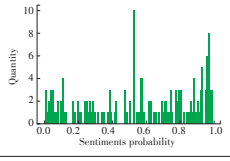
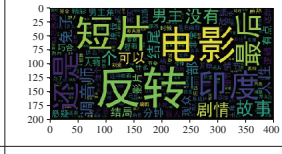
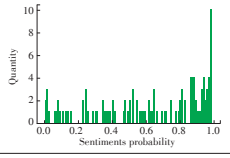
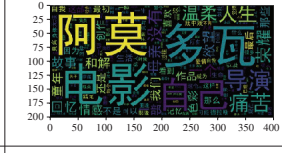
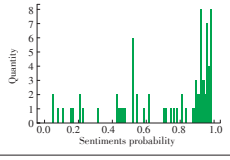
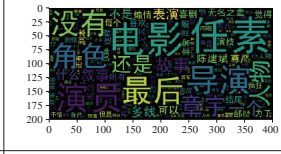
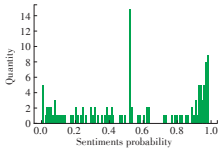
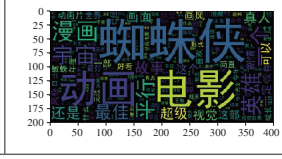
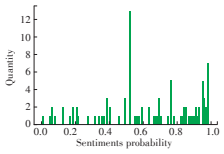
▼Table 2. Visualization analysis of experimental results from the suspense movies

Name of the movie	Visualized word cloud	Predicted sentiment probability
Fight Club		
Inception		
The Butterfly Effect		
Contratiempo		
Witness for the Prosecution		
Seven		
The Lives of Others		
Twelve Angry Men		
Infernal Affairs		
The Prestige		

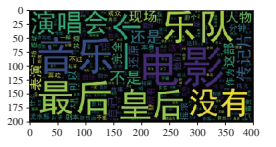
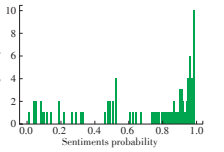
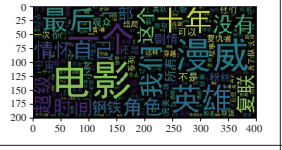
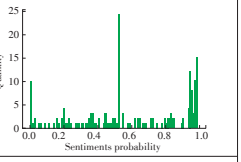
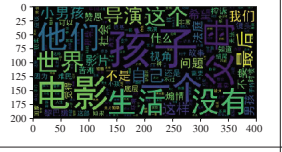
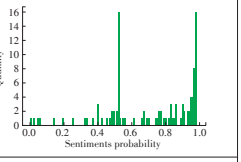
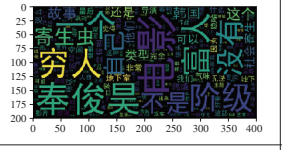
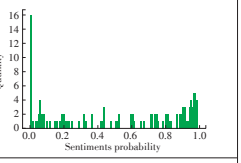
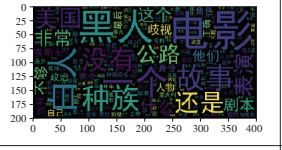
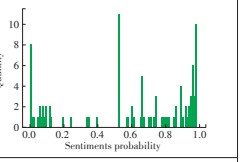
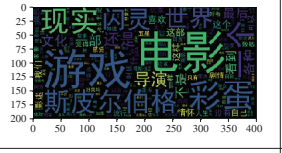
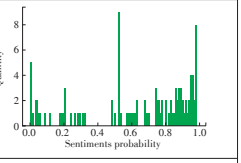
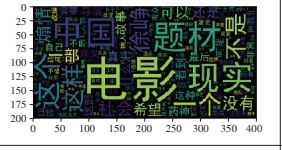
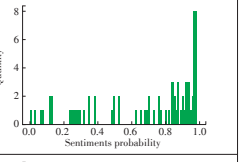
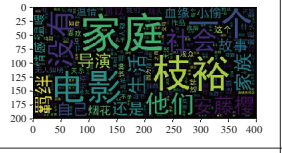
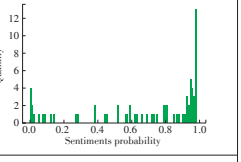
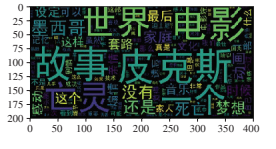
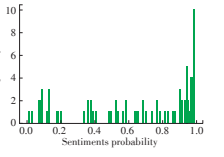
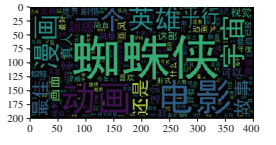
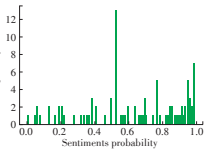
▼Table 3. Visualization analysis of experimental results from the unpopular movies

Name of the movie	Visualized word cloud	Predicted sentiment probability
The Grand Mansion Gate		
British Museum presents: Hokusai		
Dayo Wong Tze Wah Comedy Talk Show		
Zur Person: Hannah Arendt		
The Fantastic Mr Feynman		
Billy Elliot		
Curtain: Poirot's Last Case		
The Little Prince		
Daria in "Is it College Yet?"		
Daria in "Is It Fall Yet?"		

▼Table 4. Visualization analysis of experimental results from the popular movies

Name of the movie	Visualized word cloud	Predicted sentiment probability
Bohemian Rhapsody		
Avengers: Endgame		
Gisaengchung		
The Invisible Guest		
Green Book		
Hotel Mumbai		
The Blind Melody		
Pain&Glory		
A Cool Fish		
Spider-Man:Into the Spider-Verse		

▼Table 5. Visualization analysis of experimental results from the newest movies

Name of the movie	Visualized word cloud	Predicted sentiment probability
Bohemian Rhapsody		
Avengers: Endgame		
Capharnaüm		
Gisaengchung		
Green Book		
Ready Player One		
Dying to Survive		
Shoplifters		
Coco		
Spider-Man: Into the Spider-Verse		

to the Python-based sentiment analysis method, the new ideas and methods of sentiment analysis are expanded. On the other hand, our experiments still have some shortcomings; for example, the accuracy is not very high, which has a possible solution of training the class library by adding data sets to SnowNLP and using short-term emotional polarity analysis only for small project data. We hope our research has a contribution to the research of textual sentiment analysis.

References

[1] TUDORAN A A. Why do Internet Consumers Block Ads? New Evidence from Consumer Opinion Mining and Sentiment Analysis [J]. Internet Research, 2019, 29(1): 144-166. DOI: 10.1108/IntR-06-2017-0221
[2] LEE S Y, KWON Y. Twitter as a Place Where People Meet to Make Suicide Pacts [J]. Public Health, 2018, 159: 21-26. DOI: 10.1016/j.puhe.2018.03.001
[3] YANG X, XU S, WU H, BIE R. Sentiment Analysis of Weibo Comment Texts Based on Extended Vocabulary and Convolutional Neural Network [J]. Procedia Computer Science, 2019, 147: 361-368. DOI: 10.1016/j.procs.2019.01.239
[4] RASOOL A, TAO R, MARJAN K, NAVEED T. Twitter Sentiment Analysis: A Case Study for Apparel Brands [J]. Journal of Physics: Conference Series, 2019, 1176(2): 022015. DOI: 10.1088/1742-6596/1176/2/022015
[5] HUSSEIN D M E-D M. A Survey on Sentiment Analysis Challenges [J]. Journal of King Saud University-Engineering Sciences, 2016, 30(4): 330-338. DOI: 10.1016/j.jksues.2016.04.002
[6] BAGHERI H, ISLAM M J. Sentiment Analysis of Twitter Data [DB/OL]. (2017-22-25). https://arxiv.org/abs/1711.10377
[7] SAIF H, HE Y, FERNANDEZ M, et al. Contextual Semantics for Sentiment Analysis of Twitter [J]. Information Processing & Management, 2016, 52(1): 5-19. DOI: 10.1016/j.ipm.2015.01.005
[8] THELWALL M, BUCKLEY K, PALTOGLOU G. Sentiment Strength Detection for the Social Web [J]. Journal of the Association for Information Science & Technology, 2012, 63(1):163-173. DOI: 10.1002/asi.21662

Biographies

YU Qingshuang (yqs_18106301006@163.com) received the B.E. degree in Software Engineering from Qufu Normal University, China in 2019. He is currently pursuing a master's degree in computer science at China University of Petroleum (East China). His research interests include data mining and deep learning. He once participated in the National College Student Innovation Competition and won the second prize.

ZHOU Jie is currently working at the Tianjin branch of China National Offshore Oil (China) Co., Ltd. (CNOOC). After graduating in 2006, he joined the offshore oil industry and worked in China Petroleum Environmental Protection Services (Tianjin) Co., Ltd. and CNOOC Tianjin Branch. His main research interests include understanding and resolving the contradiction between the sensitive areas of the three provinces and one city, the environmental protection zone, the main functional zoning of the ocean, and the offshore oil and gas exploration and development.

GONG Wenjuan received the B.E. degree in software engineering from Shandong University, China in 2004, the M.S. degree in computer graphics from Shandong University, China in 2007, and the M.S. and Ph.D. degrees in information technology from Autonomous University of Barcelona, Spain in 2013. From 2013 to 2014, she was a postdoctoral researcher with the Oxford Brooks University, UK. She is currently with China University of Petroleum (East China). Her research interests include computer vision, audio processing, machine learning, and quantum machine learning. She has published 14 SCI-indexed papers.

From P. 01

Quality Assessment and Attribute Analysis with Multi-Task Hybrid Convolutional Neural Networks" introduces multi-task hybrid convolutional neural networks for face detection, alignment, quality assessment and attribute estimation.

The last paper "RAN Centric Data Collection for New Radio" is from the communication area, which exploits self-organizing networks and minimization of driver tests to support deployment of new radio (NR) system and conduct performance optimization.

Finally, we would like to thank all the authors, the external reviewers for their contributions and efforts to organize this special issue in this esteemed journal.