Research on News Text Clustering for International Chinese Education

Liangjie Yuan
Research Institute of International Chinese Language Education
Beijing Language and Culture University
Beijing, China
yljlarry@qq.com

Huizhou Zhao
College of Information Science
Beijing Language and Culture University
Beijing, China
zhaohuizhou@blcu.edu.cn

Zhimin Wang*
Faculty of Chinese Language and Culture
Guangdong University of Foreign Studies
Guangzhou, China
wangzm000@qq.com

***Abstract*—The automatic clustering technology for news texts is increasingly mature, enabling the monitoring and analysis of the development of international Chinese education. In this study, a large-scale corpus of news texts of international Chinese education was collected. Seven classification models were used, including logistic regression, decision trees, random forests and naive Bayes, to achieve automatic classification of news texts with an accuracy rate of 85%. Additionally, three principles for data collection and five categories for news classification were established. From the classification features of news texts, this study revealed the current development status of international Chinese education in multiple countries worldwide. Through the analysis of news from different countries, distinct characteristics were identified in terms of the content, format, and reporting style of Chinese education. These findings enable the dynamic monitoring and analysis of the development trends of international Chinese education in various countries around the world.**

Keywords—Text clustering, international Chinese education, machine learning, news texts, classifier model

I. INTRODUCTION

A. Background

In recent years, with the rise of global Chinese language learning and the increasing influence of China, international Chinese education has witnessed remarkable growth. More and more international students and scholars are choosing to learn Chinese, and Chinese companies are expanding their presence in overseas markets. This trend has fueled the rapid expansion of international Chinese education and created a significant demand in this field.

News reporting, as one of the primary channels of information dissemination, plays a crucial role in uncovering social changes and shaping public opinion. In international Chinese education, news texts hold valuable information for understanding industry dynamics and market trends. However, with the rapid development of international Chinese education, a vast amount of news texts is being generated and disseminated. The challenge lies in extracting useful information from this massive volume of text.

Consequently, research on news text clustering has become a key approach to address this issue [1]. Through clustering analysis of news texts, similar texts can be grouped together, enabling the identification of key areas and hot topics within this domain [2]. This helps educational institutions, companies, and decision-makers gain better insights into the current status and trends of international Chinese education, leading to the formulation of more scientifically sound and effective development strategies.

However, there is currently a limited amount of research focusing on news text clustering in the field of international Chinese education. Existing studies primarily concentrate on areas such as Chinese natural language processing and social media text analysis, with insufficient attention given to the specific characteristics and needs of international Chinese education. Therefore, conducting research on news text clustering specifically tailored to the development of international Chinese education holds significant significance.

B. Literature Review

The development of international Chinese education has gained significant attention in recent years, with an increasing number of individuals worldwide interested in learning the Chinese language and culture. As a result, researchers have focused on exploring various techniques to enhance the efficiency and effectiveness of delivering Chinese language education. One such area of research is news text clustering, which aims to categorize and group news articles related to Chinese language education based on their content [3].

1) Importance of News Text Clustering in International Chinese Education

News text clustering plays a vital role in the development of international Chinese education by providing a systematic way to organize and analyze large volumes of textual data related to Chinese language education. It enables educators, researchers, and policymakers to identify relevant topics, themes, and trends within the field. By leveraging clustering algorithms, it becomes possible to extract valuable insights and knowledge from diverse news sources, ultimately assisting in curriculum design, pedagogical strategies, and educational resource development.

2) Conceptual Understanding of News Text Clustering

In the realm of natural language processing, news text clustering refers to the process of categorizing or grouping similar news articles together based on their content. Clustering allows for the creation of meaningful clusters that help identify topics, themes, and trends within the news corpus. Various clustering algorithms have been applied to achieve this, such as k-means clustering, hierarchical clustering, and density-based clustering [4-5].

This work was supported by Major Program of National Social Science Foundation of China (18ZDA295); the Fundamental Research Funds for the Central Universities (19PT03); the Fundamental Research Funds for the Central Universities, and the Research Funds of Beijing Language and Culture University (23YCX155).

* Corresponding Author

3) Clustering Algorithms for News Texts

Various clustering algorithms have been explored for news text clustering. K-means clustering, as one of the most widely used methods, partitions news texts into K clusters based on similarity measures such as Euclidean distance or cosine similarity. Hierarchical clustering builds a hierarchical tree-like structure of clusters through agglomerative or divisive approaches. Density-based clustering methods, like DBSCAN, identify dense regions in the data without requiring prior knowledge of the number of clusters. Latent Dirichlet Allocation (LDA), a topic modeling technique, has also been employed to discover latent topics within news texts [6].

4) Feature Extraction Techniques

Feature extraction is a crucial step in news text clustering. Various techniques have been utilized to extract meaningful features from news texts. These include traditional approaches like bag-of-words (BoW), which represents text documents as a collection of individual words [7], and term frequency-inverse document frequency (TF-IDF), which assigns weights to words based on their importance in a document [8]. Additionally, word embeddings, such as Word2Vec and GloVe, have gained popularity for capturing semantic relationships between words [9-10]. Topic models like LDA extract latent topics from a corpus, allowing for a more comprehensive representation of news texts.

5) Applications and Challenges

News text clustering has several applications in the development of international Chinese education. It aids in organizing educational content, classifying learning resources, and identifying emerging topics and trends in the field. Additionally, it facilitates the creation of personalized learning paths and the recommendation of relevant educational materials. However, several challenges remain. Handling large-scale datasets, dealing with noisy and unstructured texts, accounting for the dynamic nature of educational topics, and addressing language-specific characteristics pose significant hurdles in implementing effective news text clustering methods for international Chinese education.

C. Research Questions and Research Methods

This study aims to address the following research questions:

1) How can news text clustering techniques be effectively utilized in the field of international Chinese education?

2) Based on news text clustering, how can we discover and analyze topics and trends related to international Chinese education?

3) What are the potentials and limitations of applying news text clustering in the field of international Chinese education?

To answer the above research questions, the following research methodology will be employed in this study:

1) By introducing the relevant concepts and literature of news text clustering in the field of international Chinese education, the current research status and important areas will be clarified.

2) Large-scale news text data related to international Chinese education will be collected, and the categories will be clustered using seven classifiers, such as Logistic Regression, Random Forest, Multinomial Naive Bayes, etc. The BERT model will be utilized to extract hot topics from the news text.

3) Through experimental data and results, the current development status and future trends of international Chinese education will be analyzed.

Using the mentioned research methods, this study will explore the potential of news text clustering in international Chinese education, providing valuable insights and suggestions for improving educational effectiveness and resource management.

II. Data Collection and Preprocessing

A. Steps of Data Collection and Processing

Data collection plays a crucial role in the research process, as it significantly impacts the quality of subsequent analysis and modeling. This study adopts a rigorous approach to data collection and preprocessing, following the steps outlined below:

1) Data Source Selection

The initial step involves carefully selecting appropriate and reliable data sources related to international Chinese education. This study focuses on gathering news text data from credible educational websites, online education platforms, and news releases by renowned educational institutions.

2) Data Acquisition

Employing advanced web scraping techniques enables us to automatically retrieve a substantial amount of news text data from the selected sources. These comprehensive news texts cover a wide range of topics and provide valuable information on Chinese education internationally. In addition, it is vital to uphold legal and ethical standards when collecting this data, ensuring compliance with relevant laws and regulations.

3) Data Cleaning

Prior to conducting any further analysis, meticulous data cleaning is performed to eliminate irrelevant or redundant information. This involves removing HTML tags, special characters, stopwords, punctuation marks, and other extraneous elements. Additionally, missing or inaccurate data are addressed to enhance data quality and consistency. Specifically, the collected news texts undergo extensive cleaning processes, such as the removal of unnecessary characters like punctuation marks, special symbols, etc. This step ensures improved text consistency and readability, facilitating subsequent processing and analysis.

4) Tokenization

Tokenization is a crucial step that transforms continuous text streams into discrete sequences of words, making them suitable for subsequent feature extraction and analysis. Popular tokenization tools like Jieba and NLTK are employed to facilitate this process. In the field of Chinese education, word segmentation is applied, dividing sentences into individual words or tokens by separating Chinese characters within the news texts. This technique aligns the data with the specific analysis requirements.

5) Stopword Removal

Following tokenization, stopwords that contribute little to text analysis and modeling are eliminated. Such stopwords include common function words, pronouns, and prepositions. Stopword removal can be accomplished using predefined lists or frequency-based methods. By removing common stop words, noise and redundancy are reduced, simplifying the subsequent analysis process. This study, for instance, eliminates frequently used vocabulary with minimal actual meaning or significance, such as "的" (*de*), "是" (*shi*) and "和" (*he*). This allows us to focus more effectively on key vocabulary and core concepts, enabling a deeper exploration of relevant issues in the field of Chinese education.

6) Text Normalization

Text normalization techniques are applied to enhance data consistency and comparability. Common practices include converting all text to lowercase, removing numbers, and eliminating symbols from words. Furthermore, techniques such as lemmatization or stemming are utilized to standardize text representation and reduce dimensionality. In the context of Chinese education, these techniques help to minimize redundancy and mitigate the impact of language variations by reducing words to their base forms or extracting word stems.

7) Data Labeling

Depending on the research objectives, news text data may be labeled for subsequent supervised learning or evaluation purposes. Examples include topic classification and sentiment analysis. Additionally, duplicate news texts are removed during the preprocessing stage to prevent bias in the analysis results. Each news text is treated as a unique entity, ensuring that it does not exert excessive influence or introduce unnecessary duplicate information.

By diligently following these data collection and preprocessing steps, clean and well-structured datasets are obtained, providing a reliable foundation for subsequent feature extraction and clustering analysis. Upholding data quality and privacy security remains a top priority throughout the process. Through comprehensive data cleaning and preprocessing of news texts in the field of international Chinese education, this study produces refined and organized data, laying a solid groundwork for further analysis and research. Consequently, it facilitates a profound understanding of trends, issues, and developmental dynamics in international Chinese education, leading to the derivation of valuable insights.

B. Data Sources and Importances

This collaborative study with listed companies has undertaken extensive data collection efforts both domestically and internationally. The collected industry data spans from 2018 to 2022 and includes key Chinese and English keywords such as "Chinese", "Chinese education", "Mandarin" and "Chinese culture". In total, these datasets amount to over 90 GB of information gathered from all around the globe.

The data sources utilized for this study encompass various platforms including the official websites of Chinese embassies abroad, Confucius Institutes, overseas Chinese schools, as well as media outlets like Chinaqw.com and Chinanews.com, among other medium. The numbers of data entries are shown in Table 1.

By adopting a comprehensive approach to international Chinese education development, the research team has extracted pertinent information pertaining to economic dimensions, industry dimensions, innovation dimensions, and other relevant aspects of industry growth.

TABLE I. COMPOSITION OF DATA

Categories	Data Sources	Numbers of data entries
Official websites	Chinese embassies	5,306
	Confucius Institutes	8,362
	overseas Chinese schools	1,215
News medium	Chinaqw.com	5,016
	Chinanew.com	6,254
	other medium	5,272

Through collaboration with the listed companies, the study has been able to acquire substantial and trustworthy data that spans critical sectors associated with Chinese education on a global scale. By meticulously organizing and analyzing this vast dataset, this study can gain valuable insights into the current state and emerging trends within the realm of international Chinese education. Consequently, these insights can provide decision-making support and strategic guidance to industries involved in this field.

Simultaneously, this multinational data collection effort aims to shed light on the holistic panorama of the international Chinese education industry. It seeks to explore the demand in Chinese education across different countries and regions. Such information holds significant importance in devising strategies for promoting Chinese language education and enhancing the international competitiveness of relevant industries.

C. Data Selection Principles

To ensure the data's representativeness, we adhere to the following data collection principles in this study:

1) Openness

All collected data were sourced from publicly available information. This means that the data came from widely accessible resources such as public databases, government statistics, news reports, official websites, and other openly accessible sources. The use of publicly available data ensures transparency and traceability, allowing other researchers to validate the research findings and conduct further analysis.

2) Richness

To comprehensively analyze and deeply understand the development of Chinese education, this study collected diverse forms of data, including both numerical and textual data. Numerical data may include statistical data, survey data, quantifiable indicators, etc., which are used to analyze and describe the quantitative characteristics and trends of Chinese education. Textual data may include academic papers, policy documents, textbooks, news reports, student essays, etc., which are used to analyze and understand the qualitative features of Chinese education, policy influences, teaching content, learning experiences, and other aspects.

3) Balance

To ensure the comprehensiveness and representativeness of the data, this study collected data from various platforms and sources. The data were not limited to specific platforms or resources but was collected from multiple platforms, including but not limited to academic databases, government websites, educational institution websites, news media, social

media, online forums, etc. By obtaining data from different platforms, biases can be reduced, and a comprehensive analysis of Chinese education can be conducted from multiple perspectives and information sources.

Additionally, data collection was conducted over a specific period of time to capture a wide range of snapshots and trends in the development of Chinese education. Multiple rounds of data verification and cross-referencing processes were employed to ensure data accuracy and reliability.

III. Classification and Clustering of News Texts on International Chinese Education

The development of international Chinese education has shown a trend towards diversification and globalization. Classifying relevant news articles can help understand the status of Chinese education in different countries and provide references. Regarding the news text data collected for this study, we can classify them into three categories as follows:

A. Classification by Countries

Due to the involvement of various countries in international Chinese education, it is necessary to classify the collected news texts by countries. This classification allows for a deeper understanding of how Chinese education is perceived and received in different regions and enables the identification of country-specific trends and perspectives.

This technology leverages algorithms and databases containing information about place names, regions, and countries to match the news texts with their corresponding locations. By analyzing contextual clues within the text such as location references, names of institutions or individuals, or specific events, geolocation technology can accurately assign news texts to the appropriate countries.

B. Classification by Categories

One important function of news texts is to report events that have already occurred. In the field of international Chinese education, news texts report on events related to Chinese education in various countries. Based on the collected news text data, we can categorize news in the field of international Chinese education into the following types:

1) Cultural Exchange Activities

This category includes cultural exchange activities related to China and Chinese culture, such as official exchanges, cultural exchanges between institutions, and summer visits. By reporting on these activities, we can observe the cooperative exchanges in Chinese education between different countries and China.

2) Courses and Examinations

This category covers information about Chinese language courses and examinations in different countries, such as introducing course information, learning content, and examination methods. These reports help understand the curriculum and examination systems of Chinese education in different countries and provide references for students and teachers.

3) Chinese Language Competitions

This category includes Chinese language competitions held in various countries, such as the famous "Chinese Bridge" competition. The content includes the name, content, participants, and awards of the competitions. Such news showcases the proficiency of Chinese learners in different countries and encourages more people to participate in Chinese language learning.

4) Personalities and Stories

This category involves special reports on students, teachers, and Confucius Institute principals who study Chinese. Through specific individuals and stories, it inspires enthusiasm and confidence in learning Chinese. These reports showcase the inspirational experiences and success stories of Chinese learners from different countries.

5) Academic and Confucius Institute Construction

This category mainly focuses on reports about Confucius Institute construction and academic conferences, reflecting the development of Chinese education institutions and Confucius Institutes in various countries. These reports demonstrate the influence of Chinese education in the academic field and provide opportunities for relevant institutions to learn from and communicate with each other.

C. Classification by News Hotspots

By considering the frequency and popularity of news events, we can obtain and rank hot news events related to international Chinese education. This classification method determines the importance and influence of news hotspots based on the authority and level of attention they receive in reporting. By paying attention to news hotspots, we can track the latest developments and trends in Chinese education and gain further understanding of the current state of Chinese education worldwide.

Also, Tracking news hotspots allows us to stay updated on the latest developments and trends in Chinese education, providing valuable insights into the current state of Chinese education worldwide. It helps us identify emerging issues, reforms, and innovations in the field, which can inform research, policy-making, and educational practices.

IV. Experimental Results and Analysis

By using the aforementioned classification methods, this study conducts three experiments to automatically classify news texts in the field of international Chinese education as follows:

A. Country-based Classification of News Texts

For the country classification of news, we employ country/city named entity recognition technology. If relevant keywords are matched, the news is labeled with the respective country tag. In this study, all news texts were classified and matched according to their respective countries. The table below shows the top 10 countries ranked by the number of news articles:

TABLE II. Top 10 countries of Country-based Classification

Rank	Country	Number of News Articles
1	UK	11,804
2	Thailand	10,508
3	USA	6,158
4	Canada	5,407
5	Malaysia	4,854
6	Japan	2,184
7	France	2,162
8	Germany	1,492
9	Russia	1,452
10	India	1,436

As is shown in Table II, we can see that: The United Kingdom (11,804 articles) ranks first, with the highest number of news articles in the field of international Chinese education. This indicates that the UK has significant influence and activities in promoting and conducting Chinese education.

Thailand (10,508 articles) ranks second, with a large number of news articles related to international Chinese education. As an important country in Southeast Asia, Thailand's emphasis on Chinese education cannot be ignored.

The United States of America (6,158 articles) follows closely behind and is also one of the countries with a relatively high number of news articles on international Chinese education. This may reflect the interest and demand for Chinese education in USA as an international academic and cultural center.

Canada (5,407 articles) and Malaysia (4,854 articles) rank fourth and fifth respectively. These two countries are also very active in investing in and developing Chinese education.

Japan (2,184 articles), France (2,162 articles), Germany (1,492 articles), Russia (1,452 articles), and India (1,436 articles) are the other countries in the top ten. The high number of news articles on Chinese education in these countries reflects their efforts and achievements in promoting and conducting Chinese education.

B. Category-based Clustering of News Texts

Following the five categories described earlier, this study proceeded by inviting experts and doctoral students specializing in international Chinese education research to annotate news texts, following the five categories mentioned earlier. Each category was annotated with approximately 500 pre-training data points. Subsequently, various classifiers were employed to automatically classify the news texts based on this annotated data.

The machine learning methods utilized in this study encompassed several statistical algorithms, including Naive Bayes, Logistic Regression, Random Forest, Decision Tree, Support Vector Machines, K-Nearest Neighbors, among others. These classifiers were applied to evaluate the accuracy of the models and determine the optimal approach for classification.

TABLE III. CONFUSION MATRIX FOR CHINESE TEXTS

No.	Model	Text Accuracy	Precision	Recall	F1
1	Logistic Regression	80.49	0.8	0.8	0.8
2	Random Forest	85.37	0.85	0.85	0.85
3	Multinomial Naive Bayes	80.49	0.8	0.8	0.8
4	Support Vector Classifer	75.61	0.76	0.76	0.76
5	Decision Tree Classifier	57.32	0.57	0.57	0.57
6	K-Nearest Neighbour	53.66	0.54	0.54	0.54
7	Gaussian Naive Bayes	41.46	0.41	0.41	0.41

The results of the classification were represented using confusion matrices, which are shown in Table III for Chinese texts and Table IV for English texts. By analyzing these matrices, we could gain insights into the strengths and weaknesses of different classifiers and assess their effectiveness in categorizing news texts accurately.

TABLE IV. CONFUSION MATRIX FOR ENGLISH TEXTS

No.	Model	Text Accuracy	Precision	Recall	F1
1	Logistic Regression	61.67	0.62	0.62	0.62
2	Random Forest	60	0.6	0.6	0.6
3	Multinomial Naive Bayes	65	0.65	0.65	0.65
4	Support Vector Classifer	60	0.6	0.6	0.6
5	Decision Tree Classifier	48.33	0.48	0.48	0.48
6	K-Nearest Neighbour	41.67	0.42	0.42	0.42
7	Gaussian Naive Bayes	31.67	0.32	0.32	0.32

From the results, we can observe that the Random Forest model achieved an accuracy rate of nearly 85% when applied to Chinese text processing. On the other hand, the Multinomial Naive Bayes model reached an accuracy rate of 65% for English text processing. These findings indicate that the Random Forest model performs better in classifying Chinese texts, while the Multinomial Naive Bayes model shows better performance in classifying English texts.

Furthermore, based on the classification results, we automatically categorized all news texts, and the final results are shown in the following figure.

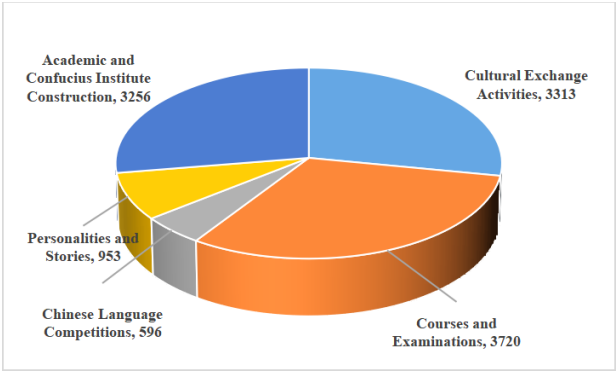

Fig. 1. The distribution of five news categories

From this, we see that in terms of news text categories, Chinese language courses and exams have the highest number, followed by cultural exchange activities, academic and Confucius Institute development, with these three categories having a similar quantity. On the other hand, there are fewer news articles related to personalities and stories, as well as Chinese language competitions. This reflects the characteristics of news content in the field of international Chinese education.

C. News Hotspot Classification

Regarding the extraction of news hot topics, we utilized the keybert and zhkeybert modules to extract key phrases using BERT embeddings. After extracting the key phrases, we ranked them based on their frequency in national statistics and selected the top 50.

Subsequently, based on the extracted keywords from each country, manual news text tracing and data re-cleaning

were conducted. By examining the original news text corresponding to the keywords, we summarized and extracted news hot topic phrases based on relevance and the content of the news text. The table below showcases the top 5 news hot topic phrases for five countries.

TABLE V. NEWS HOT TOPICS IN 5 COUNTRIES

Countries	Top 5 news hot topics
UK	Continuous development of China-UK relations
	A new chapter in China-UK cooperation
	Former Ambassador Liu Xiaoming attended the events
	Continuous advancement of China-UK friendship
	Actively contributing to the development of China-UK relations
Thailand	University-level Chinese language competition held in Thailand
	High school-level Chinese language competition held in Thailand
	Chinese language teachers in Thailand
	Promotion of Chinese culture in Thailand
	Lives and studies of Thai-Chinese individuals
USA	The United States celebrates Chinese Lunar New Year
	Updates from the Chinese Embassy and Consulates in the United States
	Traditional Chinese culture in the United States
	Chinese language teachers in the United States
	Overseas Chinese in the United States
Canada	Latest updates on Chinese schools in Canada
	Successful China-Canada International Film Festival
	Updates on China-Canada Relations
	Lives and studies of Canadian Chinese
	Promotion of Chinese Culture in Canada
Malaysia	Developments in Malaysian Chinese education
	Statement from the Malaysian Minister of Education
	Updates on Malaysian Chinese school teachers
	News from the Chinese Embassy in Malaysia
	Recent reports concerning Malaysia

From this, we can see the specific contents of Chinese education news hotspots in different countries, which further reflects the different focuses of each country in Chinese education. For example, news from the United States and the United Kingdom tends to focus more on official and leadership interactions that influence the development of Chinese education. On the other hand, Thailand and Malaysia focus on specific aspects of Chinese education such as Chinese language competitions and Chinese language teachers.

V. CONCLUSIONS

A. Key Research Findings and Contributions

This study established three principles for collecting news data and collected a large-scale corpus of international Chinese education news texts. Through techniques such as information retrieval, machine learning, automatic classification, and keyword extraction, the automatic classification of news texts was achieved. The massive news texts were classified according to three criteria: country, type, and hot topics. Furthermore, five categories were created for international Chinese education news content. Multiple machine learning models were employed as classifiers, resulting in high accuracy in the final classification results. The main findings of this study are as follows:

1) There is a higher number of reports on Chinese education in English-speaking countries and Southeast Asian countries, indicating more developed Chinese education in these regions. On the other hand, there are fewer reports on Chinese education in Africa and South America, suggesting weaker development in these regions.

2) Within the news, there is a greater emphasis on curriculum and examinations, cultural exchange activities, and the construction of Confucius Institutes, indicating learners' concerns about the Chinese courses offered by educational institutions and meeting their examination needs, while official organizations primarily report on cultural events and the development of Confucius Institutes.

3) Different countries have distinct hot topics in Chinese education news. For example, English-speaking countries tend to interpret Chinese education from a macro perspective (such as leadership visits and policy announcements), while Southeast Asian countries focus more on specific aspects of Chinese education (e.g. Chinese teachers, students, cultural activities), reflecting the differences in national contexts.

B. Limitations and Future Improvement Directions

This study has some limitations that need improvement, such as enhancing the accuracy of the classification model, refining the cleaning and annotation processes of news texts, and further improving the analysis of different countries. These aspects will be gradually refined in future work. Moving forward, we will continue to utilize advancements in scientific technology to dynamically monitor and analyze the global development of international Chinese education from news texts.

ACKNOWLEDGMENT

The authors appreciate the diligent work and helpful suggestions of reviewers. We also would like to thank Chen Zixu, Qiu Ruyi, Li Tianyi and Ma Xiaoye who participated in this research.

REFERENCES

[1] AK. Singh and M. Shashi, "Vectorization of Text Documents for Identifying Unifiable News Articles", INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, vol. 10, pp. 305–310.

[2] ZX. Fan, SY. Chen, L. Zha, JD. Yang, "A Text Clustering Approach of Chinese News Based on Neural Network Language Model", INTERNATIONAL JOURNAL OF PARALLEL PROGRAMMING, vol. 44, pp.198–206.

[3] E. Yagunova, E. Pronoza, N. Kochetkova, "Construction of Paraphrase Graphs as a Means of News Clusters Extraction", COMPUTACION Y SISTEMAS, vol. 22, pp. 1329-1336.

[4] Y. Liu and BF. Li, "Bayesian hierarchical K-means clustering", INTELLIGENT DATA ANALYSIS, vol. 24, pp. 977-992

[5] M. Vichi, C. Cavicchia, PJF. Groenen, "Hierarchical Means Clustering", JOURNAL OF CLASSIFICATION, vol. 39, pp. 553-577.

[6] L. Pion-Tonachini, S. Makeig, K. Kreutz-Delgado, "Crowd labeling latent Dirichlet allocation", KNOWLEDGE AND INFORMATION SYSTEMS, vol. 53, pp. 749-765.

[7] FX. Zeng, YF. Ji, MD. Levine, "Contextual Bag-of-Words for Robust Visual Tracking", IEEE TRANSACTIONS ON IMAGE PROCESSING, vol. 27, pp. 1433-1447.

[8] Z. Zhou, JH. Qin, XY. Xiang, Y. Tan, Q. Liu, NN. Xiong, "News Text Topic Clustering Optimized Method Based on TF-IDF Algorithm on Spark", CMC-COMPUTERS MATERIALS & CONTINUA, vol. 62, pp. 217-231.

[9] YG. Liu, ZJ. Fu, "Secure search service based on word2vec in the public cloud", INTERNATIONAL JOURNAL OF COMPUTATIONAL SCIENCE AND ENGINEERING, vol. 18, pp. 305-313.

[10] A. Mathews, SSQ. Hee, "Quantitative leak test for microholes and microtears in whole gloves and glove pieces", POLYMER TESTING, vol. 54, pp. 244-249.