Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Research on the Translation of Characteristic Words in Foreign Propaganda Translation by Data Mining Algorithms under the Big Data Platform

Junting Yang[1,] ✉

[1] *School of Foreign Studies, Wenzhou University, Wenzhou, Zhejiang, 325000, China*

ABSTRACT

This topic obtains the data of featured vocabulary under the technical architecture of big data platform and saves it in the form of dataset. Standing on the perspective of the principle of translation of featured words in foreign propaganda, the improved K-means algorithm and attention mechanism are utilized to design the translation model of featured words. The model of this paper is validated and analyzed from two aspects, namely, performance indexes and application effect, respectively. In the six performance indexes, this paper's model performs better compared to the other two control models. After the experience, the control group and the experimental group show a significant difference, i.e., the introduction of data mining algorithm is more effective in translating the featured vocabulary on the traditional model.

Keywords: k-means algorithm, attention mechanism, foreign propaganda, featured vocabulary, translation models

1. Introduction

The accumulation of China's long history and culture and the continuous development of China's society and economy have resulted in a large number of words with distinctive Chinese cultural characteristics in the Chinese language [1-2]. In today's increasingly internationalized world, there are more and more exchanges between China and the rest of the world in various aspects such as politics, economy and culture. At the same time, people all over the world pay more and more attention to Chinese culture [3-6]. Translation, as a bridge to communicate with the world, plays a very important role. How to correctly translate these words with distinctive Chinese cultural characteristics has also become the focus of people's attention more and more [7-9].

In recent years, with the wide application of databases and computer networks, together with the use of advanced automatic data generation and collection tools, the amount of data possessed by

✉ Corresponding author.
E-mail address: justin202318@126.com (J. Yang).
Received 15 January 2024; Revised 20 February 2024; Accepted 30 November 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-347

people has increased dramatically [10-11]. The contradiction between the rapid increase of data and the lagging of data analysis methods is becoming more and more prominent, people also want to be able to analyze the existing large amount of data on the basis of scientific research, business decision-making or enterprise management, but the current data analysis tools are difficult to carry out in-depth processing of data [12-15]. Data mining is precisely to solve the shortcomings of traditional analysis methods, and for the analysis and processing of large-scale data and the emergence of data mining from a large amount of data to extract the useful information hidden behind the data, it is more and more fields adopted, and achieved better results, for people's correct decision-making to provide a great deal of help [16-19], and its application in the big data platform has an important role in the translation of special vocabulary in the translation of China's foreign propaganda. Translation of Characteristic Words in Translation is of great significance.

The article explains the technical architecture of the big data platform from three aspects: data layer, computing layer and data engine, acquires network data related to foreign propaganda and featured vocabulary, and saves them in the form of data sets through data preprocessing operations. Under the dual constraints of the principle of translating featured vocabularies in foreign propaganda and the data mining algorithm, it is proposed to design the data mining-based featured vocabulary translation model by using the K-means algorithm and the Transformer architecture that integrates the attention mechanism. Performance evaluation indexes, research subjects, and test scales are determined to validate and analyze the application effect of this paper's model, respectively.

2. Research on Translation of Characteristic Lexicon under Big Data Platform

2.1. Design of the technical architecture of the big data platform

The data platform adopts open source and stable big data components to reduce research costs, and on this basis, a new integration method is proposed, namely data integration, data processing, data storage, data governance, data analysis, and data visualization.

2.1.1. Data layer. The use of a combination of different data storage modes allows for efficient and flexible management of IoT sensory data. This approach can give full play to the advantages of database and storage technologies, and provide refined management according to data characteristics and needs. The characteristics of various databases are as follows:

MySQL database is suitable for storing structured data, such as business and subject applications, because it has the characteristics of centralized data control, high independence, and good sharing, which can meet the needs of structured data.

Hadoop Distributed File System (HDFS) and FastDFS are suitable for storing large-scale file data. HDFS is suitable for cluster environments, providing highly reliable and high throughput data storage services; FastDFS focuses on file access speed and stability, suitable for scenarios with high performance requirements for file access.

Kafka and redis clusters build data buses and caching capabilities to ensure real-time data delivery and fast access. Kafka handles perceptual data and redis provides in-memory database caching and access.

MongoDB is suitable for unstructured or semi-structured data, such as metadata and device data, and its a document-based database with a flexible data model and powerful query functions.

IoT data volume is large, it is recommended to use Hbase to store cold data and openTSDB to store hot data. HBase processes massive amounts of data, and openTSDB is designed for time series data to efficiently store and query hot data.

2.1.2. Computational layer. Distributed computing service built based on Hadoop/Flink ecosystem software is an efficient and flexible data processing solution for processing massive sensory data in

the field of IoT. This solution combines the advantages of Flink and Hive, which can support real-time data computation and offline data computation services to meet different business needs, and data processing through drag-and-drop components, which is convenient.Hive SQL encapsulates the component functions and reduces the cost of learning, and at the same time, Hive is able to process the IoT sensory data in a distributed manner with the help of the MapReduce big data computing framework to Improve efficiency and scalability.Flink is a high-performance, low-latency distributed streaming data processing framework for real-time processing with high reliability and high availability.

2.1.3. Data engines. Data scheduling and task scheduling are important in big data platform architecture and determine the flow and execution of data. Using the data scheduling and task scheduling architecture, two tools, Datax and Airflow, are integrated to solve the problem of synchronization of heterogeneous data sources.Datax is a data scheduling engine that simplifies the traditional data synchronization links and enables seamless synchronization. This design improves efficiency, reduces the risk of errors, and provides strong support for data flow in big data platforms.Airflow is a task scheduling engine that solves the dependency problem between multiple tasks during big data processing. Through powerful task scheduling capabilities, it can clearly define the dependencies between tasks and automatically schedule the execution of tasks, ensuring the orderly execution of tasks and avoiding errors and delays caused by confusing task dependencies. Integrating Datax and Airflow into the big data platform architecture enables efficient and stable data processing processes. Datax is responsible for data scheduling and synchronization, and Airflow is responsible for task scheduling and execution, which improves the efficiency and accuracy of data processing and provides timely and accurate data support for enterprise decision-making.

2.2. Data Acquisition and Preprocessing

2.2.1. Data acquisition. In this paper the raw data is acquired based on a featured vocabulary database containing more than 200,000 or so foreign propaganda translations. Taking each word as a keyword for web query, traversing the featured vocabulary database by reading by line, acquiring the featured words one by one, and then using the query function of the web dictionary to replace the value of the queried keyword with the read keyword in turn to get the web page of the keyword, so as to download the web page and save it in the local disk.

2.2.2. Raw data processing. The original webpage information obtained by the theme crawler, the html file is saved in the local disk, and by observing the content and composition of its webpage data, the content extraction is carried out based on a bit of labeling, and at the same time the integration is carried out, so that the visualization results are more neat. The featured vocabulary database and English extraction results obtained in the whole process provide a good data base for future text vocabulary data analysis.

2.3. Principles of Translation of Characteristic Terms in External Publicity

Translation principles are a manifestation of translation theory and a refraction of the understanding of the essence of translation. The so-called "principle" refers to "the law or standard on which the speech or action is based", and it can be inferred that the "translation principle" refers to "the law or standard on which the translation operation is based" or "the guidelines that guide the translation operation". Translation principles are instructive for translation behavior or translation operations.

2.3.1. Differentiation and flexibility. Foreign propaganda translation is to translate a large amount of relevant information from Chinese into foreign languages accurately, rigorously and succinctly, and

to disseminate it through various channels, such as books, periodicals, newspapers, radio, television, Internet and other media, as well as through international conferences and other means, so as to satisfy the needs of foreign audiences, to set up and maintain a correct and good image of the country, and to push forward and promote the progress and development of the society as well as of the civilization of mankind. Foreign propaganda involves a wide range of contents, and different foreign propaganda has different purposes and audience needs. From the point of view of the purpose of propaganda, the purpose of different propaganda texts varies, which requires the translation to take into account the specificity of the texts in these fields. For example, when it comes to the propaganda of the country's political economy, national defense, education and other policies, the main purpose is to inform the situation, confirm the facts and disseminate the views. Through this process, the audience can get an overview of the regional development of the target country in all walks of life, and at the same time satisfy their own needs. Therefore, in this kind of translation, the translator is required to make the viewpoints clear, the arguments well-founded, convince people with reasons, take a clear position, try to make it easy for others to accept, and at the same time, pay special attention to the accuracy of the language. For some tourism materials, the purpose of publicity and the needs of the audience are different from the former. The purpose of publicity is mainly to attract more foreign tourists, enhance their love for our country's natural landscape, national customs, and further promote the tourism industry. Therefore, this kind of foreign publicity text will need to strengthen the content of subjectivity, characteristics and knowledge, and the language should be concise, popular and easy to understand. From the above two cases, we can see that when translating foreign propaganda, different discourse processing strategies should be adopted according to the different contents involved in the propaganda materials, different styles should be used and different languages should be used.

2.3.2. The “inside-out” principle. Usually, all materials to be publicized abroad need to be close to the reality of China's development and the needs of foreign audiences. For those who are engaged in the work of translating foreign communications, the most important thing to pay attention to is that they should study foreign cultures and foreigners' psychological thinking patterns, be good at discovering and analyzing the subtle differences between Chinese and foreign cultures and their respective characteristics, and be good at grasping the inherent logic of the two different cultures and languages as well as the differences in their expressions. We should be good at discovering and analyzing the subtle differences between Chinese and foreign cultures and their respective characteristics, at grasping the inner logic of two different cultures and languages and the differences in their expressions, and at all times grasping the translation in accordance with the habits of thought and language of the foreign audience, so as to achieve the best dissemination effect. It should be noted that in the context of cultural globalization, foreign propaganda translation should also pay attention to “differences between the outside and the outside”, the so-called “differences between the outside and the outside” refers to the fact that different countries have different languages and cultures, and their audiences' appreciations and needs are also different, so it is only natural that audiences outside of China are also different. All audiences outside of China are different, of course. Translation should be carried out according to the needs of different audiences in different cultural backgrounds, so as to make the translation of foreign propaganda more targeted, close and timely, and to maximize the communication effect.

2.3.3. Principle of cultural adaptation. Cultural translation adaptability means that in translation practice, the expression of cultural information should be adapted to the cultural reality and development needs of the target language: from the point of view of the value standard of ideology and conception, any target-language culture tends to be positively accepted and assimilated, and it will be difficult to realize the effect of acceptance and assimilation if the form of cultural expression is out of step with the cultural reality of the nation. From the perspective of linguistic and cultural

reality, the development of language has stages, and the linguistic characteristics of different historical periods often have different acceptance standards; from the perspective of readers, any reader is a historical person and a real person, and every reader has a certain experience of cultural history and cultural reality. The so-called "cultural adaptation" does not mean that the target language culture should be adapted to the culture of the original language, but that each culture has its own heterogeneity, so there are certain limits of translatability in the actual translation process. The purpose of adhering to the principle of cultural adaptation in translation practice is to make the target language come into contact with foreign cultures from all aspects through cultural translation, and to rise to a new culture that is more vital and more adaptable to new historical and ecological conditions by absorbing the essence of foreign cultures. In this sense, following the principle of cultural adaptability has practical meaning and important guiding significance for foreign propaganda translation under the background of cultural globalization.

2.4. Data Mining Based Translation Model for Featured Terms

In this subsection, under the guidance of the big data platform and the principle of featured vocabulary translation in foreign propaganda, it is proposed to adopt the K-means algorithm and the Transformer architecture that incorporates the attention mechanism to jointly construct a data mining-based featured vocabulary translation model, and the detailed structure is as follows:

2.4.1. Improved K-means algorithm. The specific flow of the ICKM algorithm is shown in Fig. 1.In data mining technology, K-means is a commonly used method, which has a very good application in data sets with obvious features [20-21]. The study uses the ICKM algorithm to determine the initial center point, and the original set of sample points is reclassified into K sets as a way to keep the error sum of squares of the algorithm stable. The distance between data objects can be expressed by equation (1).

$$d_{i,j} = ((X_i - X_i)^T S^{-1}(X_i - X_i))^{\frac{1}{2}} \tag{1}$$

In Equation (1), d_{ij} represents the distance between data objects and S represents the sample covariance matrix. The algorithm selects any point in the space as the centroid and makes a circle with the distance r as the radius and the region in the circle is called the neighborhood of the centroid.

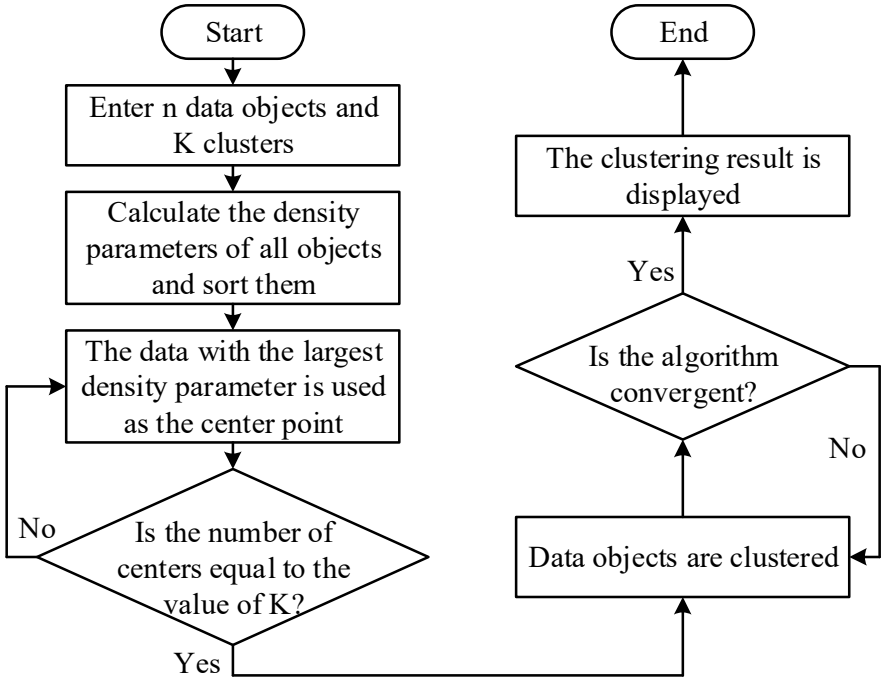


Fig. 1. ICKM algorithm flow chart

2.4.2. Transformer architecture incorporating attention mechanisms. The structure of decoder-encoder translation is shown in Fig. 2. Neural machine translation is a commonly used machine translation model, which is based on the traditional statistical translation model for nonlinear mapping of neural networks, so as to realize bilingual conversion [22-23]. When in the translation system, the information obtained using data mining technology needs to be coded and decoded mechanism to get accurate translation.

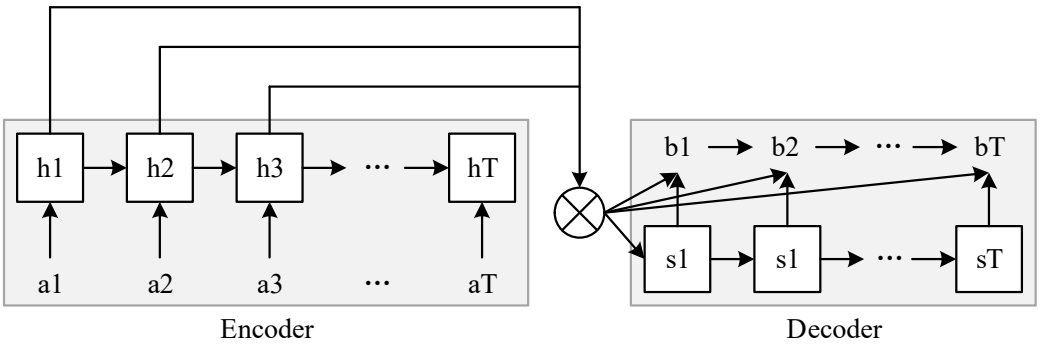


Fig. 2. Decoder-encoder translation structure

Figure 2's a is the sample with translation, b is the translation result, h is the hidden state at the moment of encoding, and s is the hidden state at the moment of decoding. When the translation is in progress, the neural network needs to calculate the hidden vector of the i th word at the moment of t, which can be calculated by the hidden state of the previous moment and the transformation function, as shown in equation (2):

$$h^t = f(a^t, h^{t-1}) \tag{2}$$

The state transition through Eq. (2) will be followed by the conversion of the hidden state to a context vector using a nonlinear activation function as in Eq. (3):

$$w = F(h^1 \cdots h^t) \tag{3}$$

Where F is the conversion function and the terminated h^t is the final result of the context vector. On the basis of this calculation result the conversion function g is used to calculate the hidden state in the decoding process, which is calculated as in Eq. (4). That is:

$$s^t = g(b^{t-1}, w, s^{t-1}) \tag{4}$$

The generation probability can be calculated by softmax function. The maximum probability of sequence decoding can be estimated by similarity maximization in the training process, which is implemented as in equation (5). That is:

$$\begin{aligned} &P(b^1, \cdots, b^{i-1} \mid a^1, \cdots, a^i) \\ &= \prod_{i=1}^{i} P(b^i \mid b^1, \cdots, b^{i-1}, a^1, \cdots, a^i) \\ &= \prod_{i=1}^{i} P(b^i \mid b^1, \cdots, b^{i-1}, w) \end{aligned} \tag{5}$$

What also gives the output sequence is the loss function, as shown in equation (6). Namely:

$$\begin{aligned} &-\log P(b^1, \cdots, b^{t-1} \mid a^1, \cdots, a^t) \\ &= -\sum_{i=1}^{i} \log P(b^i \mid b^1, \cdots, b^{t-1} w) \end{aligned} \tag{6}$$

The loss function in the formula can evaluate the advantages and disadvantages of the translation effect, the smaller the loss function, the better the translation effect, therefore, the translation model is optimized by reducing the value of the loss function during the operation of the translator. This combination of decoder and encoder translation method defect is easy to forget the earliest encoded sentence vectors, weighting results will be added to the input text sequence, decoding translation can be based on the vector set by the attention mechanism for key information extraction, at this time, the decoder will be changed, i moment of the hidden state s^i is calculated as in equation (7). That is:

$$s^i = f(s^i - 1, b^{i-1}, w^i) \tag{7}$$

Where w^i is the context vector, which aims to influence the weight of a certain information in the generated word, and is calculated as in equation (8). Namely:

$$w^i = \sum_{j=1}^{r} \alpha^{ij} h^j \tag{8}$$

where h^j is the hidden vector of the j nd word in the decoding process, and α^{ij} is the weight between word i and word j in the translation generation, and the formula is shown in equation (9):

$$\alpha^{ij} = \frac{\exp(d^{ij})}{\sum_{k=1}^{i} \exp(d^{ik})} \tag{9}$$

Where d^{ij} is the degree of influence of word i and word j, and k is the total number of generated words. After the experience of the attention mechanism, the level of neural machine translation will be effectively improved.

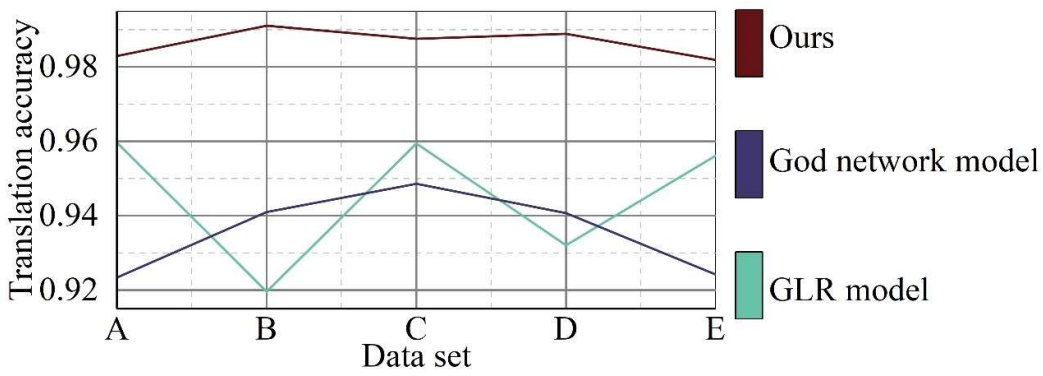
3. Example Analysis of Featured Lexical Translation Model

3.1. Model validation analysis

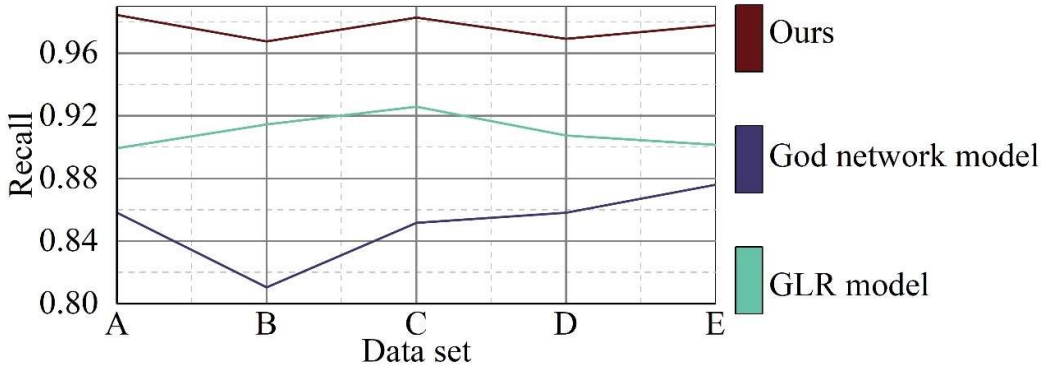
3.1.1. Data sets. In order to test the effectiveness of the paper's method in improving the performance of machine featured vocabulary translation, comparative experiments are conducted on the basis of the above. A, B, C, D and E featured vocabularies are selected as the test data sets for the model under study, each of the five featured vocabularies contains 1388, 5867, 106, 18557 and 9647 articles, and each featured vocabulary library contains 58,265, 85,645, 2,847, 5,221,138, and 2,154,590 sentence pairs, respectively. The featured vocabularies contain six different types of articles: sports, military, economy, education, science and technology, and society.

3.1.2. Evaluation indicators. The research dataset is determined above, and the evaluation indexes and control model are set next. The GLR model and the neural network model are selected as the comparison models, and translation precision, semantic information recall, feature matching of subject words, F1 value, BLEU value, and GLEU value are selected as the evaluation indexes to measure the performance of machine English translation. The feature matching degree of subject words is an important index to measure the degree of matching between the source language and the target language, and the BLEU value is a measurable case-sensitive index applied to the evaluation of the translation effect, which is calculated by using multi-bleu.perl script to calculate the BLEU value of the translation result, and the higher the BLEU value, the better the translation effect of the machine English translation model, and the GLEU value is a variant of BLEU, which is the index of the evaluation of machine translation, and the GLEU value is a variant of BLEU. The GLEU value is a variant of the indicator BLEU, which is often applied in machine translation evaluation, and the GLEU value can effectively measure the fluency of the utterance after machine translation.

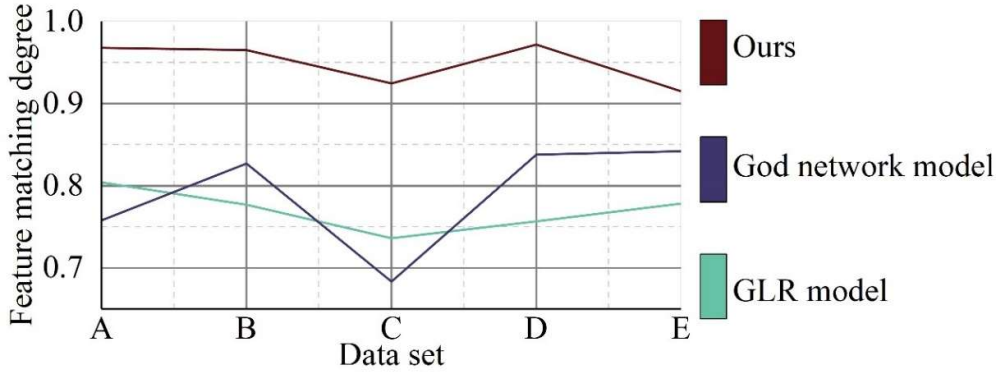
3.1.3. Data analysis. The performance comparison results of translating different featured lexicons using the three models are shown in Fig. 3, where (a) to (f) are translation precision, semantic information recall, feature matching of subject words, F1 value, BLEU value, and GLEU value, respectively. Analyzing the experimental comparison results in Fig. 3, it can be seen that in terms of translation precision, the translation precision of translating statements within different featured vocabularies using the model of the paper is higher than 98.23%, while the translation precision of translating statements within different featured vocabularies based on the GLR model as well as the neural network model is lower than 96.13%, and it is judged that the translation precision of the model of the paper for translating statements within different featured vocabularies is significantly higher than that of the other two models It is judged that the translation precision of this model is significantly higher than the other two models. The same situation exists in the five indexes of semantic information recall, feature matching of subject words, F1 value, BLEU value and GLEU value. Overall, the model test verifies that the model has better translation effect of featured vocabulary, and the semantic information obtained from the translation results has high recall and high feature matching degree of subject words, which proves that the model under study has high translation accuracy and rationality, and it has a promoting effect on the translation of international cultural propaganda.



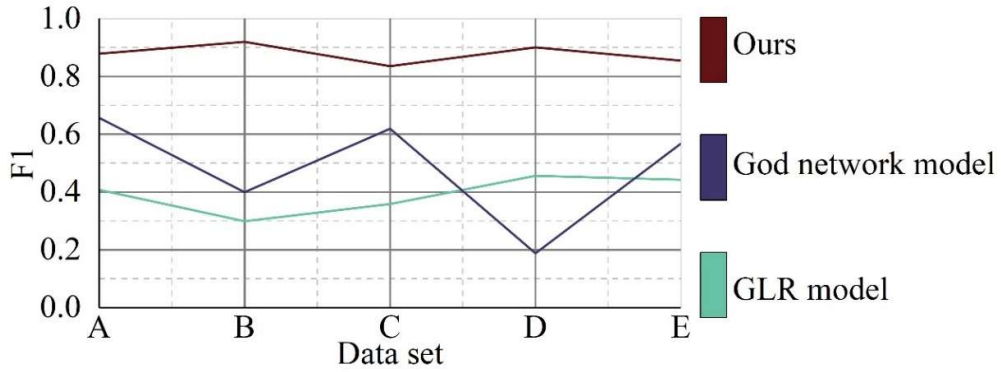
(a)Translation accuracy



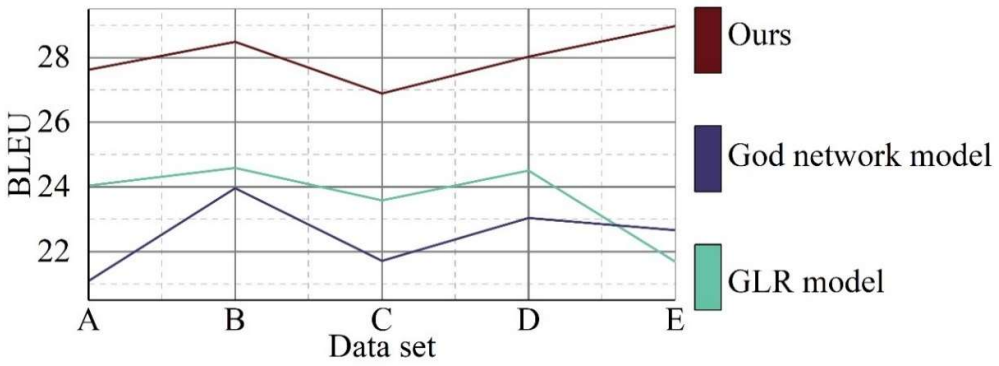
(b)Recall



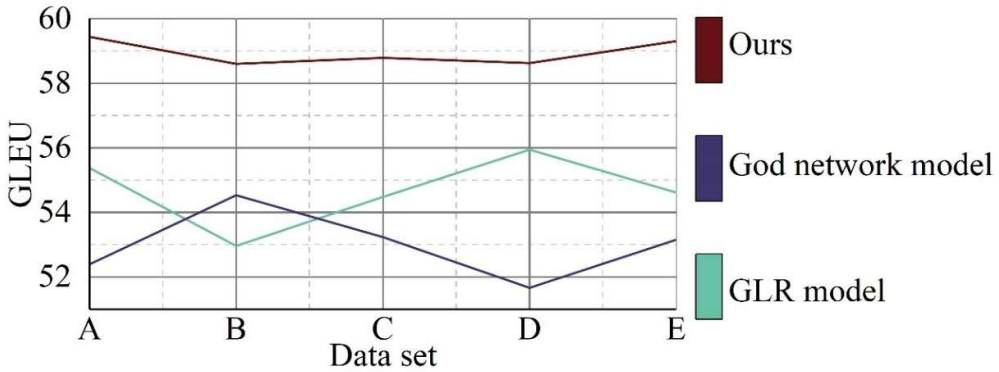
(c)Feature matching degree



(d)F1-Value



(e)BLEU



(f)GLEU

Fig. 3. Performance comparison result

In order to further test the semantic coverage of the research method, based on the above experimental environment, the evaluation results of testing the online generation of featured vocabulary translations are derived as shown in Table 1.The higher the semantic coverage, the higher the matching of featured vocabulary translations, and the better the effect of online generation of featured vocabulary translations. Analyzing Table 1, it can be seen that the semantic coverage of the five featured terminology libraries tested using the research method is higher than 85%, with the highest reaching 96.97%. It indicates that the stability of the research's featured vocabulary translation online generation control is better, and the featured vocabulary has a high degree of matching.

Table 1. Semantic coverage evaluation results

Characteristic vocabulary set	Number of semantic concepts	Correct relational quantity	Semantic coverage
A	132	128	0.9697
B	166	157	0.9458
C	137	118	0.8613
D	157	144	0.9172
E	138	121	0.8768

3.2. Analysis of the effect of practical application of the model

3.2.1. Practical application programs. The rapid development of big data technology is profoundly affecting the translation industry, and at the same time bringing new opportunities and challenges for the development of foreign propaganda translation. As a breakthrough technology of big data platform, the machine model based on data mining shows great potential in featured vocabulary translation. For example, with the technical support of the big data platform, highly simulated publicity translation environments can be created, such as international conferences, business negotiations, contract signing and other scenes. Allowing subjects to practice featured vocabulary

translation in these virtual scenarios allows them to more intuitively feel and cope with the complexity and challenges of the actual work of foreign propaganda translation. This kind of immersive practice experience can help the subjects better understand the differences in characteristic vocabulary, business etiquette and the use of industry terminology in the translation process. In the following section, we will verify and analyze the effectiveness of the model's practical application scheme.

3.2.2. Study subjects. The experimental subjects are 60 subjects, 30 in the control group and 30 in the experimental group, and all 60 belong to professionals in the direction of featured vocabulary research. The control group (traditional translation mode) and the experimental group (fusion data mining translation mode) consisted of 15 male and 15 female subjects each, in order to avoid the influence of gender factor on the results of the experiment as much as possible, which also ensured the reliability and validity of the experiment to a certain extent. The scale was conducted before the subjects did not experience the two translation modes to verify whether the sample met the requirements of the study. After the homogeneity test, the subjects were allowed to experience the two modes separately, and the quantitative data were obtained from the scale test for the subjects after the experience, and the independent samples t-test was conducted on the quantitative data to confirm the practical efficacy of the model in this paper.

3.2.3. Quantitative data collection. Designing the Test Scale for Accuracy of Featured Terminology Translation, which is divided into three parts, the first part is accuracy of featured terminology translation, which consists of 20 items. The second part is Specialty of Featured Terminology Translation, which consists of 30 items. The third part is featured vocabulary translation completeness, containing 25 items. The experimental and control groups were tested using a 5-point scale scoring method, which has high reliability and validity performance. After completing the data collection process, it was uniformly saved in the spreadsheet Excel and the scale data were analyzed as independent samples using SPSS26.0 software.

3.2.4. Analysis of results:

1) Pre-experience analysis. In order to test whether there is a significant difference between the selected research samples (also known as the homogeneity test), in this regard, the SPSS software was used to conduct an independent samples t-test on the scale test data of the two groups before the experience, and the results of the pre-experience analysis are shown in Table 2. From the data in Table 2, it can be seen that there is no significant difference between the experimental group and the control group in the featured vocabulary translation dimensions (accuracy, completeness, and professionalism) before the experience, i.e., $P>0.05$, which indicates that the selected research samples meet the experimental requirements.

Table 2. Analysis results before intervention

Dimension	Experimental group(N=30)		Control group(N=30)		T-Value	P-Value
	Mean	Standard deviation	Mean	Standard deviation		
Accuracy	2.46	0.35	2.42	0.33	-2.331	0.144
Integrity	2.44	0.36	2.41	0.29	-2.791	0.135
speciality	2.39	0.38	2.36	0.35	-2.119	0.108

2) Post-experience analysis. After the research samples passed the homogeneity test, the data of the two groups' scales were next analyzed for differences after experience, and the results of the analysis of differences between the two groups after experience are shown in Table 3. As can be seen through the data in the table, after the subjects' experience, it was found that the experimental group and the control group produced significant differences in the featured vocabulary translation dimensions (accuracy, completeness, and professionalism), which satisfies the condition of $P<0.05$. That is to say,

it indicates that adding data mining algorithms to the traditional model helps to improve the quality of foreign propaganda translation.

Table 3. Analysis results of difference between the two groups after intervention

Dimension	Experimental group(N=30)		Control group(N=30)		T-Value	P-Value
	Mean	Standard deviation	Mean	Standard deviation		
Accuracy	3.57	0.44	2.57	0.38	2.131	0.007
Integrity	3.51	0.48	2.49	0.41	2.299	0.001
Speciality	3.52	0.46	2.62	0.42	1.487	0.005

4. Conclusion

In this paper, with the technical support of the big data platform, the data studied in this paper are obtained, and the original data are preprocessed. Combined with the principle of translating featured words in foreign propaganda, it is proposed to use the improved K-means algorithm and translator to jointly construct a featured word translation model, and the model is used to carry out an example study and analysis. In the six indexes, the model of this paper has a higher priority than the other two models, which proves that the translation model has a good performance. In addition, after a period of experience, the experimental group and the control group produced significant differences in the dimensions of featured vocabulary translation (accuracy, completeness, and professionalism), indicating that the featured vocabulary translation model incorporating data mining algorithms has a particularly objective effect of enhancing the development of publicity translation.

Funding

No funding were used to support this study.

Data Availability Statement

The datasets generated during and/or analyzed during the current study are not publicly available due to sensitivity and data use agreement.

Conflicts of Interest

The authors declare that there is no conflict of interest with any financial organizations regarding the material reported in this manuscript.

References

[1] Qi, L., Wang, Y., Chen, J., Liao, M., & Zhang, J. (2021). Culture under complex perspective: a classification for traditional chinese cultural elements based on nlp and complex networks. Complexity, 2021(1), 6693753.

[2] Yu, J., Jian, X., Xin, H., & Song, Y. (2017, September). Joint embeddings of chinese words, characters, and fine-grained subcharacter components. In Proceedings of the 2017 conference on empirical methods in natural language processing (pp. 286-291).

[3] Wang, L. (2014). Internationalization with Chinese characteristics: The changing discourse of internationalization in China. Chinese Education & Society, 47(1), 7-26.

[4] Xiong, Q. (2023). International Exchanges and Cultural Communication in Chinese History. Lecture Notes on History, 5(1), 63-69.

[5] Shuyu, L. (2024). The significance and prospects of world civilization exchanges. Academic Journal of Humanities & Social Sciences, 7(1), 56-61.

[6] Mulinda, C. K. (2015). On Cultural and Academic Exchanges between China and African Countries. International Journal of Asian Social Science, 5(4), 245-256.

[7] House, J. (2020). Translation as a prime player in intercultural communication. Applied linguistics, 41(1), 10-29.

[8] Hamaidia, L., Methven, S., & Woodin, J. (2018). Translation spaces: Parallel shifts in translation and intercultural communication studies and their significance for the international development field. Translation spaces, 7(1), 119-142.

[9] Chromá, M. (2014). The role of translation in professional communication. The Routledge Handbook of Language and Professional Communication, 147-164.

[10] Richa, M., Prévotet, J. C., Dardaillon, M., Mroué, M., & Samhat, A. E. (2021, November). An automated and centralized data generation and acquisition system. In 2021 28th IEEE International Conference on Electronics, Circuits, and Systems (ICECS) (pp. 1-4). IEEE.

[11] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. Information systems, 47, 98-115.

[12] Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, 3(1), 91-99.

[13] Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. Array, 16, 100258.

[14] Li, B., Hou, Y., & Che, W. (2022). Data augmentation approaches in natural language processing: A survey. Ai Open, 3, 71-90.

[15] Islam, M. (2020). Data analysis: types, process, methods, techniques and tools. International Journal on Data Science and Technology, 6(1), 10-15.

[16] Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.

[17] Gan, W., Lin, J. C. W., Chao, H. C., & Zhan, J. (2017). Data mining in distributed environment: a survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(6), e1216.

[18] Olson, D. L., Lauhoff, G., Olson, D. L., & Lauhoff, G. (2019). Descriptive data mining (pp. 129-130). Springer Singapore.

[19] Galdi, P., & Tagliaferri, R. (2018). Data mining: accuracy and error measures for classification and prediction. Encyclopedia of bioinformatics and computational biology, 1, 431-436.

[20] Vassilios Yfantis,Achim Wagner & Martin Ruskowski. (2025). Federated K-means clustering via dual decomposition-based distributed optimization. Franklin Open100204-100204.

[21] Hatice Nur Karakavak,Hatice Oncel Cekim,Gamze Ozel & Senem Tekin. (2025). Seismic Microzonation and Earthquake Forecasting in Western Anatolia Using K-Means Clustering and Entropy-Based Magnitude Volatility Analysis. Geotechnical and Geological Engineering(2),99-99.

[22] Yuting Zuo,Jing Chen,Kaixing Wang,Qi Lin & Huanqiang Zeng. (2025). Local flow propagation and global multi-scale dilated Transformer for video inpainting. Journal of Visual Communication and Image Representation104380-104380.

[23] M. Srati,A. Oulmelk,L. Afraites,A. Hadri,M.A. Zaky & A.S. Hendy. (2025). On a computational paradigm for a class of fractional order direct and inverse problems in terms of physics-informed neural networks with the attention mechanism. Journal of Computational Science102514-102514.