International Journal of Web-Based Learning and Teaching Technologies
Volume 19 • Issue 1

Utilizing Natural Language Processing to Enhance Ideological Education in Tibetan Universities

Quan Yang, GuangXi Vocational and Technical Institute of Industry, China
Huajian Xin, GuangXi Vocational and Technical Institute of Industry, China*
Xuehua Ji, GuangXi Shangmao Senior Technical School, China
Fae Mai, Western Iowa Tech Community College, USA

ABSTRACT

This article analyses the significance of ideological and political education for college students in Tibet and proposes a natural language model for an ideological education curriculum to improve the accuracy of students' document search. The segmentation results are optimized to enhance literature search accuracy, promoting the development of ideological education in Tibetan universities. By utilizing N-gram language models and enhanced technical modes, key information can be quickly obtained, developing students' interest in ideological education. The construction of a corpus is deemed crucial for expedited access to ideological education documents. The study suggests that combining information technology with ideological education can create new opportunities for innovation and reform in the field.

KEYWORDS

college students, mind education, natural language processing, Tibetan schools, unsupervised learning

INTRODUCTION

Constrained by factors like poor natural conditions and a weak economic foundation, Tibet's transportation infrastructure has long lagged behind (Wang.T et al., 2020). Before liberation, there was no modern highway, and the transportation infrastructure relied on people and horses. The vast agricultural and pastoral areas were, in essence, a relatively closed environment. Since the democratic reform, great improvements in both the central government's support and the transportation infrastructure in and out of Tibet have been achieved (Gao et al., 2020).

The continuous improvement of airport facilities has led to a modernized transportation network, with road transportation as the mainstay and railway and air transportation serving as supplements. This transformation has greatly strengthened the connection between Tibet and the mainland (Mukherjee, 2021). Tibetan college students who grew up in this evolving environment exhibit strong personalities

DOI: 10.4018/IJWLTT.337390 *Corresponding Author

and resilience to hardships and hard work. On the other hand, due to the relatively closed living environment and limited exposure to information, their perspectives tend to be narrow, and their thinking relatively conservative (Lhagyal, 2021).

With the increasingly urgent situation in China's anti-secession struggle and the pervasive effects of the information society, it is obvious that traditional explicit ideological and political education is unable to meet the needs of contemporary education goals. In the information society, college students, including those in Tibet, are susceptible to encountering bad or illegal information, making them open to its influence. Therefore, there is an urgent need for Tibetan colleges and universities to strengthen ideological and political education to navigate the complexities of the information age.

The ideological and political education of ethnic minorities is an important component of China's developmental strategy for ideological and political education. In Tibet, an area with unique historial, cultural, and ethnic characteristics, ideological and political education is a crucial task for college students (Yang & Leibold, 2020). Tibet has been an integral part of China since ancient times, with its development and stability closely related to national security and unity. As such, the cultivation of patriotic and law-abiding college students it is essential, as they can contribute to the social and economic progress of Tibet and the country (Zhang & Tsung, 2020).

Still, ideological and political education in Tibetan universities face many challenges, including the influence of separatist forces, ineffective teaching methods, and low levels of information technology (Yeh & Makley, 2019). To address these issues, this article proposes a natural language model in the ideological education curriculum. This model aims to improve the accuracy of students' document searches and enhance their interest and motivation to learn ideological and political knowledge. Recognizing students as important talents in national development, it emphasizes the value of aligning with the trends of the information society. Students are encouraged to actively improve their proficiency in information technology and enhance their awareness of information search. The emergence of document retrieval, propelled by advancements in information technology, highlights the goal of teaching: to improve students' information literacy.

Natural language processing (NLP) is a branch of artificial intelligence (AI) that deals with the interaction between human languages and computers. It can analyze, understand, generate, and manipulate natural language texts or speech (Bum, 2018). NLP finds applications in a variety of fields, such as information retrieval, machine translation, sentiment analysis, text summarization, and question answering (Foggin, 2018). This article aims to explore the use of NLP technology to strengthen ideological and political education in Tibetan universities.

The significance of ideological and political education for Tibetan college students cannot be ignored, yet they are currently facing various challenges. In response to these issues, this study proposes a natural language model for the ideological education curriculum, which comprises three primary components: (1) segmentation optimization; (2) literature search enhancement; and (3) corpus construction. This article analyzes the importance and challenges of ideological and political education of Tibetan college students, introduces a natural language model for ideological education courses, and uses NLP technology to improve the accuracy and efficiency of literature retrieval related to ideological and political education in Tibetan colleges and universities.

MATERIALS AND METHODS

Literature Review

Information retrieval refers to the organization and search of information resources, such as literature. Its main purpose is to meet the specific needs of users by employing search methods to retrieve all relevant literature while discarding documents that appear relevant but do not meet the specified requirements.

Some researchers have proposed using NLP technology to improve the document search and retrieval process. For example, Wang Y. et al. (2020) proposed a Tibetan-Chinese bilingual document retrieval system based on word embedding and cross-language query expansion. They used a pretrained word embedding model to represent Tibetan and Chinese words as vectors, applying a cross-language query expansion methods to enrich query terms with synonyms and related words from both languages. They evaluated their system on a Tibetan-Chinese bilingual document dataset, demonstrating its superiority over several baselines in terms of accuracy and recall.

Some researchers suggest using NLP technology to generate personalized content for students. For example, Li et al. (2019) proposed a natural language generation system for ideological education based on reinforcement learning and attention mechanisms. Their system optimizes content generation through reinforcement learning models based on student feedback and uses attention mechanisms to focus on relevant information in input text. Evaluation on an ideological education text dataset revealed that their system produced more coherent, diverse, and informative text than several baselines.

Furthermore, other researchers suggested NLP techniques for assessing and improving students' understanding and writing skills. Zhang et al. (2018) proposed a natural language understanding system for ideological education based on deep neural networks and knowledge maps. Their approach uses a deep neural network model to extract semantic features from students' answers to ideological education questions and employing a knowledge graph model to represent background knowledge and logical relationships of ideological education concepts. Evaluation on an ideological education question and answer dataset a high degree of accuracy and consistency in grading students' answers.

The theoretical framework of this study revolves around using NLP technology to strengthen ideological and political education in Tibetan universities. While many related studies exist in the literature, most of the research focuses on employing NLP technology for text classification, sentiment analysis, and text summarization. There has been relatively little exploration of its application in the field of ideological and political education.

Although some studies combine information technology with ideological and political education, most are limited to AI-based assisted teaching and student behavior analysis, lacking analysis of the content of ideological and political education. In addition, existing research has also paid little attention to the construction of a corpus suitable for the field of ideological and political education to better support literature retrieval and knowledge acquisition.

Therefore, this study aims to use existing NLP technology, combined with the characteristics of ideological and political education in Tibetan universities, to build a corpus suitable for this field. The goal is to improve the quality and efficiency of ideological and political education through improved literature retrieval and knowledge acquisition processes.

Construction for Corpus

Segmentation optimization is the process of dividing a text into meaningful units, such as words or phrases (Wangdu, 2020). The task is fundamental in NLP as it affects the performance of subsequent tasks, such as parsing, tagging, or indexing (Wang et al., 2021). However, segmentation is not a trivial problem, especially for languages with complex morphology or a lack of clear word boundaries, such as seen in Tibetan. Therefore, this study proposes the use of an N-gram language model to optimize the segmentation results of Tibetan texts.

An N-gram is a sequence of N consecutive words or characters in a text. An N-gram language model is a probabilistic model that estimates the likelihood of an N-gram based on its frequency in a large corpus. Using an N-gram language model enhances the accuracy and efficiency of Tibetan text segmentation.

Literature search enhancement is a process aimed at improving the quality and relevance of document retrieval for students' ideological education (Zhang & Tsung, 2019). Literature search is an essential skill for college students, helping them acquire new knowledge, develop critical thinking

skills, and write academic papers (Gros, 2019). However, literature search can be challenging for students unfamiliar with the topic or search tools. Therefore, this study proposes the use of NLP techniques to enhance literature search for ideological education in Tibetan universities (Zhang et al., 2018). Specifically, the study uses keyword extraction, query expansion, document summarization, and document ranking to help students in finding and accessing relevant documents more easily.

Keyword extraction is the process of identifying the most important words or phrases in a text that represent its main topic or content. This can assist students in formulating effective queries for literature search. Query expansion is a process of adding more terms to a query to increase its coverage and recall, enabling students to retrieve more documents related to their query, even if they do not contain the exact query terms.

Document summarization, a process of producing a concise and informative summary that captures the main points or highlights of a document (Huaguo, 2018), can help students quickly browse and understand the content without reading it in full. Document ranking involves ordering a set of documents according to their relevance or importance to a query. This helps students prioritize and select the most useful documents for their learning.

Corpus construction is the process of collecting and organizing a large set of texts relevant to a specific domain or topic (Greitens et al., 2019). A corpus is a valuable resource for NLP research and applications by providing linguistic data and statistical information for various tasks. This article proposes the construction of a corpus tailored for ideological education in Tibetan universities. It will contain texts from various sources, including textbooks, journals, newspapers, and websites, covering topics like socialism with Chinese characteristics, the Chinese dream, Tibetan history and culture, national unity, and security. By constructing such a corpus, the study aims to provide swift access to ideological education documents for students and teachers.

In the ideological and political education of Tibetan college students, the construction of a corpus is of great practical significance for facilitating rapid access to ideological and political education documents (Gyal, 2019). Teachers of ideological and political education theory courses should adapt their teaching methods, tailoring their lecture content according to varying classroom conditions and audiences. Furthermore, they should closely combine ideological and political education courses with information technology. The integration of wearable system and corpus technology provides a novel approach for the reform and innovation of ideological and political education theory courses.

Words, being the smallest, independent, and meaningful language components, serve as the basis for typical information retrieval systems establish indexes based on each independent word. These systems output documents based on the location and frequency of words during queries. When constructing a corpus for ideological and political education literature, the main vocabulary can be divided into three aspects: (1) professional vocabulary; (2) hot vocabulary; and (3) commonly used vocabulary (Ptackova, 2019).

Professional vocabulary includes words like ideological overview, historical events, and policy discussions, which have strong domain characteristics and obvious features of domain word formation (Bodenhorn et al., 2020). Hot vocabulary mainly includes new policies, network terms, slogan strategies, and popular events, with variations across different periods. Common vocabulary includes everyday phrases and common word collocations (Zhao et al., 2021). This article focuses on these three aspects of vocabulary, considering them as the primary components of a corpus (Li et al., 2013).

However, the specific ratio of these three types of vocabulary in the corpus should be planned carefully to ensure an optimal language model training result. Given the technical and specialized nature of ideological and political education, an inappropriate ratio could lead to some vocabulary dominating the entire model training process, shifting the focus and skewing the resulting language model (Wang et al., 2017). Therefore, it is essential to consider the composition ratio of the corpus when obtaining language materials.

Word Sample

The character-based word segmentation scheme (N-gram segmentation) is a purely mechanical method for word segmentation. This technique cuts a document into consecutive 1 metacharacter, 2 metacharacter, or multiple-characters units. Notably, the requirements for the training set differ. The word-level N-gram training set is trained independently, while the word-level N-gram training set is based on another. Such a training set for the language model does not require labelling, but it requires a large-scale corpus for model training. Additionally, the corpus must undergo a standardized process before use.

The standardized process involves separating the text in the corpus by newlines and ensuring the words are separated by spaces. This confirms that the text in the corpus is in units of words, which is convenient for subsequent training.

The word-level language model treats words as individual units. The word-level language model can predict the nth word according to the first (n-1) words. In a sequence M composed of n words, the calculation formula for the word-level N-gram language model is:

$$P(\left|w_{n1}, w_{n2}, \ldots w_{n-1}\right.) \tag{1}$$

In the formula, w1, w2, …, wn-1 is the (n-l) word in the sentence and wn is the nth word per the conditional probability formula:

$$P\frac{B}{A} = \frac{P(A,B)}{P(A)} \tag{2}$$

The segmentation path probability of sequence M can be obtained as:

$$P(w_{n1}w_{n2}\ldots w_n) = P(w_1) * P(w_2\left|w_1\right.) * \ldots * P(w_n\left|w_1\ w_2 \ldots w_{n-1}\right.) \tag{3}$$

The length of the sequence M is not fixed. Thus, when faced with a short Chinese text sequence, equation (3) can be used to obtain the corresponding result faster. It is corrected as:

$$P(w_{n1}w_{n2}\ldots w_n) = P(w_1) * P(w_2\left|w_1\right.) * \ldots * P(w_n\left|w_{n-1}\right.) \tag{4}$$

where P(wnlWn-1)

$$\frac{Count(w_n, w_{n-1})}{Count(w_{n-1})} \tag{5}$$

Training the Language Model

Language models are often confused with word embedding. The main difference is in the significance of word order. In a language model, word order is very important because it attempts to capture the context between words. On the other hand, in word embedding, only semantic similarity is captured because it is trained by predicting the words within a window, regardless of the order (Wang et al., 2018). After obtaining a language model for the field of ideological and political education, it can

be further trained according ideological and political education. The specific process of N-gram language model training is shown in Figure 1.

Through the statistical analysis of the N-gram model, the accuracy of word segmentation has been improved, and certain stop words have been removed. This results in more meaningful subsequent operations and minimizes the computational burden. The final set of all word items processed by the N-gram method is marked as N. Each word in N has the potential to become a keyword and must participate in the subsequent calculation of criticality weights.

The text in the corpus is separated by newlines, and words are separated by spaces. This labelling approach aims to establish a training set with words as a unit, replacing the traditional manual labelling training set. The test set is mainly constructed through self-labelling documents related to ideological and political education. The effectiveness of the final word segmentation can be verified through the test set.

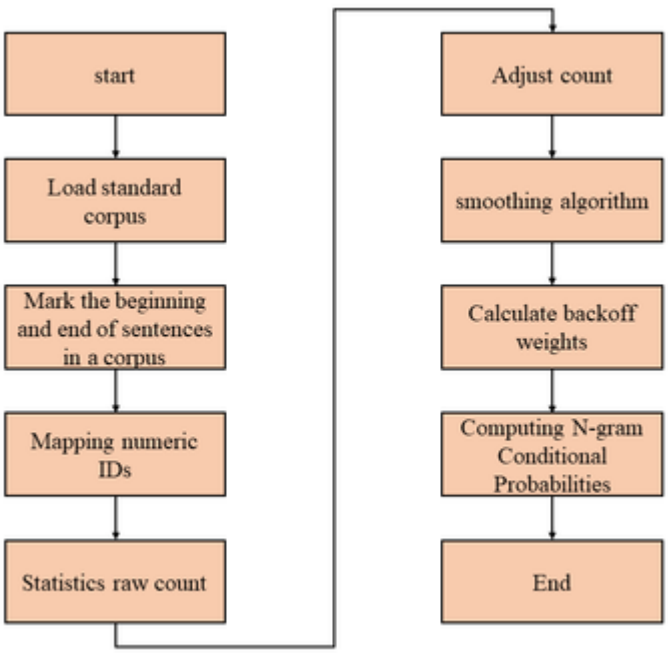
Calculate the Word Segmentation Path

There are significant differences in the word segmentation methods between Chinese and English. In English word segmentation, one word represents one word. The Chinese word segmentation rule uses Chinese characters as the writing unit. There is no clear distinction mark between words, requiring manual methods for segmentation. For a text with a sequence length of M to be segmented, its path probability is:

$$P = P(s_1)P(s_2)P(s_3)\ldots P(s_l) \tag{6}$$

The Viterbi algorithm, one of the widely used dynamic programming algorithms, is mainly used to identify the Viterbi path—the hidden state sequence that is most likely to generate the observed event sequence. This algorithm is widely used in CDMA and GSM digital cellular networks, dial-up

Figure 1. Flow chart of N-Gram language model training



modems, satellites, deep space communications, and 802.11 wireless networks. Nowadays, it is also often used in speech recognition, keyword recognition, computational linguistics, and bioinformatics.

For example, in speech recognition, where the sound signal is regarded as the observed sequence of events, and the text string is seen as the implicit cause of the sound signal, the Viterbi algorithm can be applied to the sound signal to find the most likely text string. It is also employed to find the optimal word segmentation path for a sequence, particularly in the case of Chinese text where there can be multiple possible word segmentation results in the sequence. Assuming x represents the jth possible value of the state ci and expanding the state sequence, the optimal path graph structure of the Viterbi algorithm can be obtained (see Figure 2).

Considering that the model finally adopts a four-gram language model, the traditional word segmentation problem can be solved as a labelling problem. Words are labelled words at different positions in a single word, and calculations are performed based on these labelled words. Here, "b" is assumed to represent a single-character word or the first character of a multi-character word, "c" represents the second character, "d" represents the third character, and "e" represents the rest of the multi-character word. For a word ck in the sentence, the calculation can be expressed as:

$$P(b) = P(c_k) \tag{7}$$

$$P(c) = P(c_k \left| c_{k-1}\right.) \tag{8}$$

$$P(d) = P(c_k \left| c_{k-2}c_{k-1}\right.) \tag{9}$$

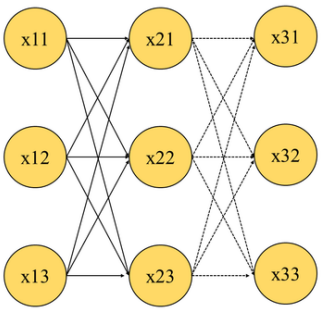
$$P(e) = P(c_k \left| c_{k-3}c_{k-2}c_{k-1}\right.) \tag{10}$$

The probability in the formula can be obtained from the trained word-level language model. Non-zero transition probabilities include P(b|b), P(d|b), P(b|c), P(d |c), P(b|d), P(e|d), P(b|e), and P(e|e). By setting of the transition probabilities, the algorithm directly determines the long word and final word segmentation result. The frequency of occurrence of short words, according to the transition probability, yields the optimal word segmentation path—resulting in the optimal word segmentation result.

Through the analysis of the word segmentation results of more samples, it can be found that traditional word segmentation tools ultimately yield unsatisfactory results for specialized words with special word formation methods. Specialized words within a specialized corpus conform to two distinct characteristics: local aggregation and global sparseness.

Local aggregation refers to the composition of related professional vocabulary that often differs from ordinary vocabulary, showing strong professionalism compared with common vocabulary. These words serve as key vocabulary in related field literature, appearing in a large number of professional documents related to the research direction. On the other hand, global sparsity refers to the particularity

Figure 2. Optimal path graph structure



of professional vocabulary, mainly appearing in a large number of professional documents related to less related research directions. However, their frequency of occurrence in other documents is often extremely low, resulting in an overall lower frequency in the corpus.

Considering these two characteristics, the concept of word frequency deviation is introduced.

$$TED(t) = \frac{1}{m-1}\sum_{i=1}^{m}(TF(t,d_i) - \overline{TF}(t,D))^2 \tag{11}$$

The calculation of the ranking word frequency deviation is shown in equation 12.

$$rTED(t) = RANK(t)^{-\alpha}\left|t\right|^{\beta} \tag{12}$$

This article uses the ranking word frequency deviation index as the word segmentation index and optimizes word segmentation for the aforementioned unsupervised Chinese word segmentation algorithm. The score is output as the final word segmentation result, aiming to achieve word segmentation optimization. First, the candidate word segmentation result set is defined as follows:

$$\begin{cases} t_I \oplus \ldots \oplus t_I = w_I \oplus \ldots \oplus w_K \\ t_j = w_i \oplus \ldots \oplus w_{i+\Delta} \\ I \le j \le l, l \le i \le K, O \le \Delta \le K \end{cases} \tag{13}$$

The word segmentation score of a candidate word segmentation result "s" of a sentence is calculated by the product of the corresponding sorted word frequency deviations of each word segmentation unit. The specific formula is shown in equation 14.

$$SCORE(\vec{s}) = \prod_{j=1}^{l} rTFD(t_j) \tag{14}$$

The word combination can be formalized as the following optimization problem (see equation 15):

$$\arg\max SCORE(\vec{s}) \tag{15}$$

Experiments show that due to the exponential size of POST(s), considering practical system applications, the extremely high amount of computation and time consumption pose significant challenges. Therefore, this article transforms the problem into solving the sequence decoding problem. By combining the concept of dynamic programming, an efficient solution based on the Viterbi algorithm is designed and implemented. The final segmentation result is obtained through this calculation.

RESULT ANALYSIS AND DISCUSSION

As an important aspect of ideological and political education, college students should improve the political and ideological education curriculum in schools and provide them with sufficient political

and ideological learning resources. However, given the literature on ideological and political education and the diverse interest of individuals, students can benefit from searching the literature independently to cultivate their specific interests in political and ideological education.

Regarding the corpus of commonly used words, this section of the corpus was mainly obtained from about 100,000 articles from People's Daily, utilizing crawlers and other means. Among the commonly used Chinese word segmentation corpus, the most representative and influential is the People's Daily word segmentation corpus from Peking University. In the development of this corpus, annotation standards were proposed and retrieval methods were studied. The proportion of the corpus construction is shown in Figure 3.

In today's Chinese word segmentation system, there are still two urgent issues that need to be addressed. The first issue is ambiguity, as many ambiguous phenomena persist in Chinese. Specifically, it refers to the occurrence of multiple different results for a few word segmentations. The second issue is the recognition of unregistered words. The system should be able to recognize and record words that have not yet been logged in. Through continuous organization, it should further enhance the recognition ability of words that have not been logged in.

The professional corpus is the largest in the political and ideological education corpus. Considering the characteristics of long vocabulary and complex structure in the field of ideological and political education, this article uses the ranking order, namely the four-gram language model, when selecting the number of "It," to verify the advantages of the four-gram language model and other sequential language models in the field of Chinese document segmentation.

N-gram language models suggest that the higher the order of the model, the longer the time required to process the Chinese word segmentation task. Through research and practice, it is found that the order of high-level language models like five-gram is greater than four, resulting in an increased training time and training set demands, leading to a significant increase in segmentation time. The one-gram language model has a special application scenario. In the application of word segmentation, its performance is relatively poor with insufficient recognition ability. Therefore, only two-gram, three-gram, and four-gram are analyzed here.

The test set is classified according to the number of words in the text, which includes three types of texts to be segmented: less than 5,000 words; greater than or equal to 5,000 words but less than or equal to 10,000 words; and greater than 10,000 words. The specific comparison test results are shown in Figure 4.

Figure 4 illustrates that the abscissa represents the type, and the ordinate axis is the processing time. The figure demonstrates that with the increase in the order of the language model and the number of words of the text to be segmented, the processing time for Chinese word segmentation also increases. However, the time difference is relatively small. At the same time, in the comparison

Figure 3. Corpus construction scale diagram

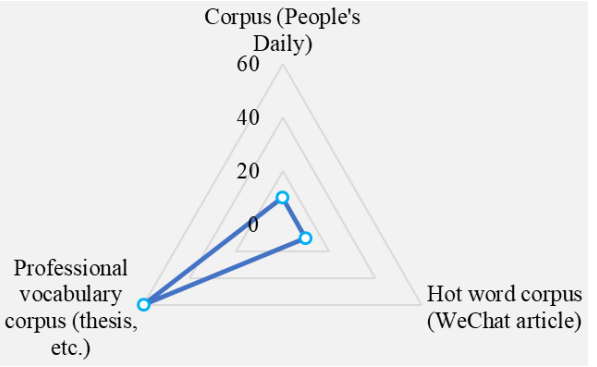
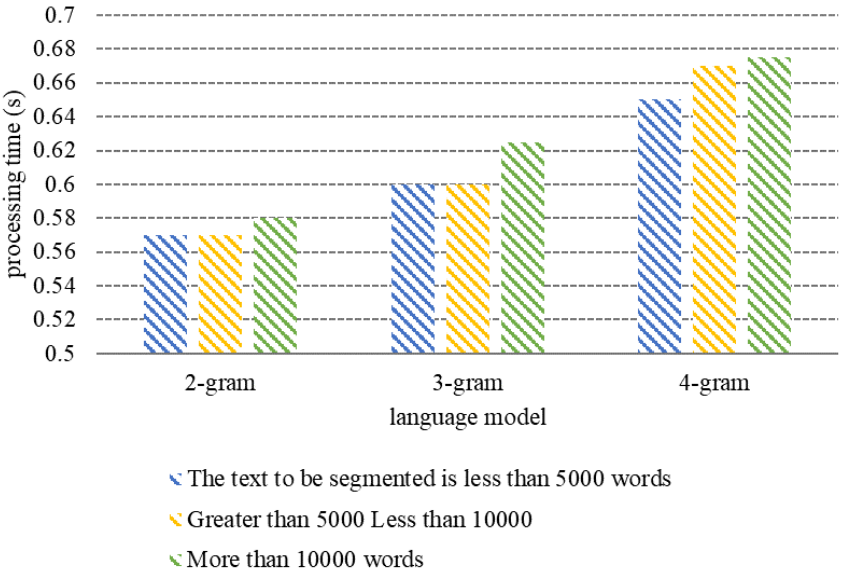


Figure 4. Comparison of processing time with different orders



of the processing time of the three types of language models, the change of the language model order does not greatly affect the processing time for different numbers of words. The time difference in the processing time for the text to be segmented remains relatively stable, showing that the algorithm has practical applications within the system. Thus, the use of the four-gram language model is practical.

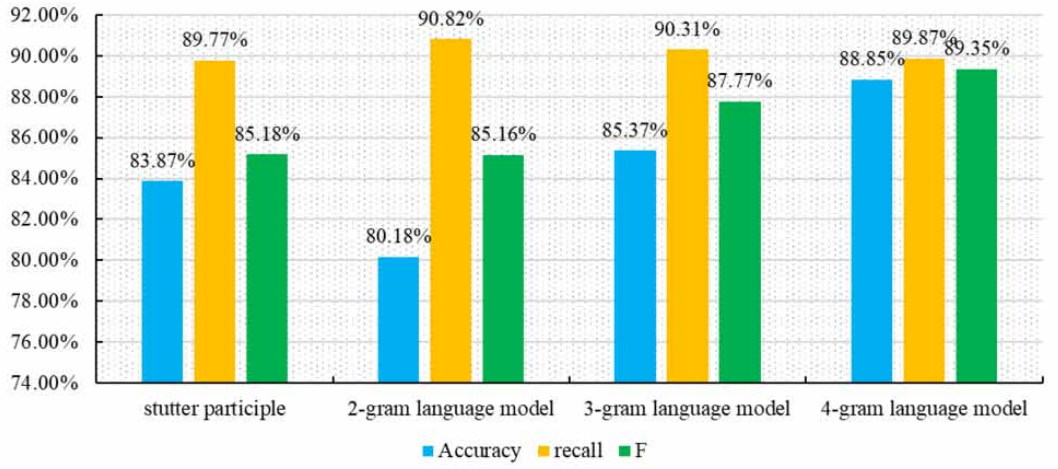
Next, the focus shifts to the word formation characteristics of domain vocabulary and testing literature that meets such requirements. For Tibetan college students in plateau areas, the use of the N-gram language model can quickly obtain keywords, saving students time in searching for literature. As text information resources increase, increasing the efficiency of natural language models for text retrieval becomes necessary. Improving the technical model also aligns ideological and political education courses more closely with modern teaching practice, which not only benefits students but promotes the development of ideological and political education courses.

To test the accuracy of natural language model searching for ideological education courses, this study conducts a word segmentation test for stuttering word segmentation using three language models, each evaluated based on three index values. The results are then calculated and statistically analyzed. The corresponding index values are obtained, and the comparison results are shown in Figure 5.

According to the comparison of word segmentation results, the experimental results can be summarized into the following aspects:

1. The accuracy is only higher than that of the Chinese word segmentation algorithm based on the two-gram language model. The difference can be attributed to the method of word segmentation, and the corpus built in the previous article can better aid the model in improving the recognition ability of professional vocabulary.
2. It can be seen from the indicators, such as the accuracy of the Chinese word segmentation algorithm implemented based on language models of different orders, that the algorithm is more inclined to segment longer vocabulary. This inclination is particularly evident in the four-gram. After the corpus is further improved, there is potential for improvement in its accuracy and other indicators.

Figure 5. Word segmentation result graph

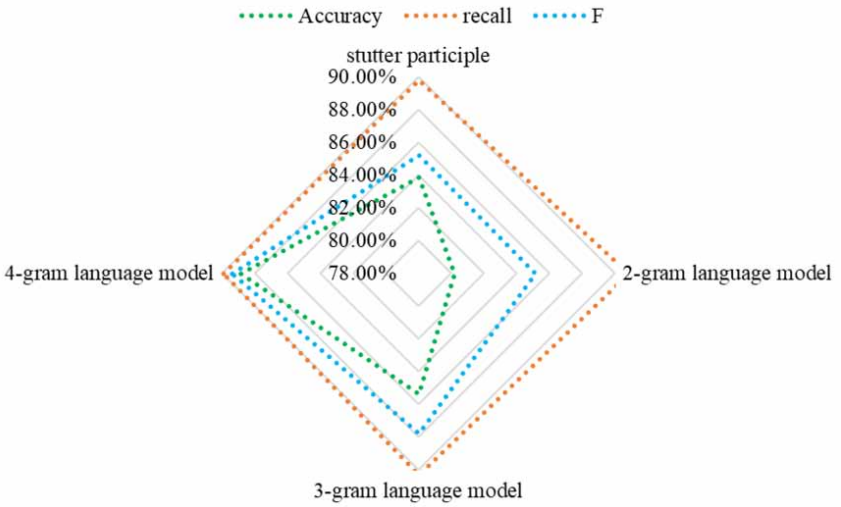


3. Considering the comparison of the processing time for different orders in the previous article, the disadvantage of the four-gram language model in processing time is not obvious. Also, the four-gram language model has a greater advantage in accuracy. However, the experimental results reveal that the four-gram language model still has room for improvement in the segmentation of some prepositions, a key area for optimization in ideological and political education.

To verify the effect of the Chinese word segmentation optimization algorithm on Chinese word segmentation, this article uses multiple word segmentation algorithms to generate preliminary word segmentation results. It then optimizes these preliminary results and compares the outcomes. The specific results are shown in Figure 6.

According to the comparison of different algorithms in the optimized word segmentation result table, it is evident that the Chinese word segmentation optimization algorithm has certain advantages

Figure 6. Optimization word segmentation result table



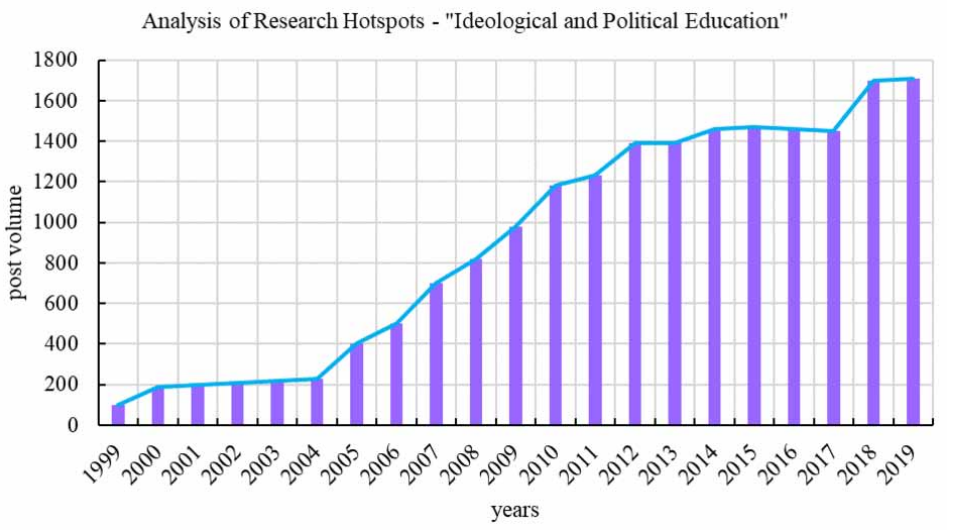
in the optimization of word segmentation results. The accuracy of multiple algorithms has been improved to a certain extent. The selected four-gram language model in this article has contributed to an improvement in the index values of the Chinese word segmentation algorithm. However, in the optimization of the Chinese word segmentation algorithm based on the two-gram language model, the enhanced accuracy rate has led to a significant drop in the recall rate. The F value has also decreased. This may be attributed to the algorithm separating too many single-character or double-character words. During the optimization process of word reorganization, this leads to numerous wrong reorganizations, ultimately leading to unsatisfactory results.

In conclusion, the Chinese word segmentation optimization algorithm has a positive impact on Chinese word segmentation. It proves to be a feasible approach to optimize Chinese word segmentation in the field of ideological and political education through vocabulary reorganization. The optimized word segmentation results also improve accuracy in literature searches related to ideological and political education, allowing Tibetan college students in plateau areas to swiftly access the desired knowledge. Given that educational literature is the focus of their learning process, improving the accuracy of literature searches fosters an environment conducive to students' exploration of ideological and political education, thereby increasing their enthusiasm.

The research hotspot analysis function compiles statistics on research hotspots according to different keywords, drawing relevant histograms to reflect the number of publications related to a specific keyword. This feature is convenient for users tracking research hotspots, comprehending the development trends, and gaining insights into the status of research hotspots. The diagram is shown in Figure 7.

Figure 7 illustrates the word segmentation results of research hotspots under the term "ideological and political education." The vertical axis represents the number of articles published (in units of articles), while the horizontal axis denotes the year of publication, displayed according to relevant information icons. The changes in the number of published papers in the past 20 years are displayed, offering insight into the changing trend of publication numbers. The research hotspot map mainly provides users with the ability to grasp the development and changes of research hotspots within the word segmentation results. For Tibetan college students, this resource allows them to quickly focus on the latest knowledge in political and ideological education literature, ensuring they stay current

Figure 7. Research hotspot map



and update their concepts. This is conducive to Tibetan college students in increasing their awareness of maintaining national unity.

Strengthening the ideological and political education of college students in Tibet is a long-term and arduous task. First, to enhance ideological and political education, there is a need to reinforce the teaching of theoretical courses, invest in the development of teaching staff, reform teaching methods and means, and constantly improve teaching quality and effectiveness. Second, making full use of new media is crucial, enabling timely and targeted ideological and political education initiatives. Third, publicizing exemplary models of college students and showcasing the positive ideological and political energy of excellent teachers and students through public initiatives is essential.

He et al. (2021) found various problems in ideological and political education that need to be improved and strengthened. The internet holds a strong attraction for college students, but many lack a clear purpose when going online. Despite the crucial role ideological and political education plays in improving teaching quality by helping college students establish correct learning attitudes, values, life perspectives, and social attitudes, the professionalism of teachers often lacks precision and effectiveness in ideological and political education.

In response to the importance and challenges of ideological and political education of Tibetan college students, this article puts forth methods to strengthen ideological and political education in Tibetan colleges using NLP. Specifically, it employs natural language models, word segmentation optimization, literature search enhancement, and corpus construction for ideological and political education courses to improve the accuracy and efficiency of student literature retrieval. The experimental results show that this algorithm can quickly obtain keywords for Tibetan college students, saving them time in literature searches. By combining information technology with ideological education, new opportunities can be created for innovation and reform in this field.

Strengthening ideological and political education requires not only optimizing teaching methods and building a teaching staff but also utilizing new media and promoting outstanding college student models. This approach fully leverages the influence and appeal of the positive ideological and political energy of excellent teachers and students.

CONCLUSION

Due to its special geographical environment, unique historical and cultural context, and complex political environment, Tibetan colleges and universities show great differences in knowledge base, thought processes, and competition consciousness between Tibetan college students and mainland college students. Ideological and political education in Tibetan colleges and universities bears greater responsibility for shaping students' ideological and political thoughts compared to mainland institutions.

This study explores the application of Chinese word segmentation optimization algorithms in the field of ideological and political education through the analysis of the corpus. Experimental results show that the four-gram language model has significant advantages in accuracy, and the Chinese word segmentation optimization algorithm effectively improves the accuracy of word segmentation. Overall, the word segmentation algorithm used in this article has certain advantages compared with traditional word segmentation tools in the field of ideological and political education. Still, further improvements to the corpus and optimization of the algorithm for preposition segmentation can improve its performance. In the future, studies should aim to refine the corpus and optimize algorithms to better serve Tibetan college students and their social environment.

DATA AVAILABILITY

The figures used to support the findings of this study are included in the article.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING STATEMENT

This work was supported by Key Project of Guangxi Vocational Education Teaching Reform Research in 2021 by the Education Department of Guangxi Autonomous Region (Gui Jiao Zhi Cheng [2021] No.34)- "Research on Talent Training Path of Intelligent Manufacturing Specialty Group under the Background of Integration of Industry and Education" (No. GXGZJG2021A008).

ACKNOWLEDGEMENT

The authors are appreciative of those techniques that have contributed to this research.

REFERENCES

Bodenhorn, T., Burns, J. P., & Palmer, M. (2020). Change, contradiction and the state: Higher education in greater China. *The China Quarterly*, *244*, 903–919. doi:10.1017/S0305741020001228

Bum, T. (2018). Translating ecological migration policy: A conjunctural analysis of Tibetan pastoralist resettlement in China. *Critical Asian Studies*, *50*(4), 518–536. doi:10.1080/14672715.2018.1515028

Foggin, J. M. (2018). Environmental conservation in the Tibetan Plateau region: Lessons for China's Belt and Road Initiative in the mountains of Central Asia. *Land (Basel)*, *7*(2), 52. doi:10.3390/land7020052

Gao, S., Li, S., & Zhang, Q. (2022). The creative route of university ideological and political teaching on the essential of embedded sensor network. *Wireless Communications and Mobile Computing*, *2022*, 1–12. doi:10.1155/2022/4004415

Greitens, S. C., Lee, M., & Yazici, E. (2019). Counterterrorism and preventive repression: China's changing strategy in Xinjiang. *International Security*, *44*(3), 9–47. doi:10.1162/isec_a_00368

Gros, S. (2019). China virtual and real: "Minzu" spaces. *Cross-Currents: East Asian History and Culture Review, 1*(32).

Gyal, H. (2019). "I am concerned with the future of my children": The Project Economy and shifting views of education in a Tibetan pastoral community. *Critical Asian Studies*, *51*(1), 12–30. doi:10.1080/14672715.2018.1544500

He, X., Dong, X., Liu, L., & Zou, Y. (2021). Challenges of college students' ideological and political and psychological education in the information age. *Frontiers in Psychology*, *12*, 707973. doi:10.3389/fpsyg.2021.707973 PMID:34484065

Huaguo, Z. H. A. N. G. (2018). Analysis on development of green economy in Tibet from the perspective of ecological civilization. *Journal of Landscape Research*, *10*(5), 62–78.

Lhagyal, D. (2021). 'Linguistic authority' in state-society interaction: Cultural politics of Tibetan education in China. *Discourse (Abingdon)*, *42*(3), 353–367. doi:10.1080/01596306.2019.1648239

Li, J., Liu, Y., Wang, H., & Liang, X. (2019). Natural language generation for ideological education based on reinforcement learning and attention mechanism. *Jisuanji Yingyong*, *39*(11), 3137–3142.

Li, X. L., Gao, J., Brierley, G., Qiao, Y. M., Zhang, J., & Yang, Y. W. (2013). Rangeland degradation on the Qinghai-Tibet plateau: Implications for rehabilitation. *Land Degradation & Development*, *24*(1), 72–80. doi:10.1002/ldr.1108

Mukherjee, K. (2021). Conflict in Tibet: Internal and external dimensions. *Asian Affairs*, *52*(2), 288–311. doi:10.1080/03068374.2021.1907103

Ptackova, J. (2019). Traditionalization as a response to state-induced development in rural Tibetan areas of Qinghai, PRC. *Central Asian Survey*, *38*(3), 417–431. doi:10.1080/02634937.2019.1635990

Wang, P., Wang, X., Wang, C., Miao, L., Hou, J., & Yuan, Q. (2017). Shift in bacterioplankton diversity and structure: Influence of anthropogenic disturbances along the Yarlung Tsangpo River on the Tibetan Plateau, China. *Scientific Reports*, *7*(1), 12529. doi:10.1038/s41598-017-12893-4 PMID:28970506

Wang, T., Yan, J., Cheng, X., & Yu, Y. (2020). Irrigation influencing farmers' perceptions of temperature and precipitation: A comparative study of two regions of the Tibetan Plateau. *Sustainability (Basel)*, *12*(19), 8164. doi:10.3390/su12198164

Wang, W., Zhao, X., Li, H., & Zhang, Q. (2021). Will social capital affect farmers' choices of climate change adaptation strategies? Evidences from rural households in the Qinghai-Tibetan Plateau, China. *Journal of Rural Studies*, *83*, 127–137. doi:10.1016/j.jrurstud.2021.02.006

Wang, Y., Zhang, Y., Liang, J., & Liang, X. (2020). Tibetan-Chinese bilingual document retrieval based on word embedding and cross-lingual query expansion. *Jisuanji Yingyong*, *40*(9), 2605–2610.

Wangdu, K. (2020). Decoloniality, power and ideology in the social studies textbooks of Tibetan exile schools. *Journal of Curriculum Studies*, *52*(2), 230–246. doi:10.1080/00220272.2019.1657958

Yang, M., & Leibold, J. (2020). Building a "double first-class university" on China's Qing-Zang Plateau: Opportunities, strategies and challenges. *The China Quarterly*, *244*, 1140–1159. doi:10.1017/S030574102000106X

Yeh, E. T., & Makley, C. (2019). Urbanization, education, and the politics of space on the Tibetan Plateau. *Critical Asian Studies*, *51*(1), 1–11. doi:10.1080/14672715.2018.1555484

Zhang, C. L., Li, Q., Shen, Y. P., Zhou, N., Wang, X. S., Li, J., & Jia, W. R. (2018). Monitoring of aeolian desertification on the Qinghai-Tibet Plateau from the 1970s to 2015 using Landsat images. *The Science of the Total Environment*, *619*, 1648–1659. doi:10.1016/j.scitotenv.2017.10.137 PMID:29061294

Zhang, L., Tsung, L., & Zhuoma, . (2020). Exploring sustainable multilingual language policy in minority higher education in China: A case study of the Tibetan language. *Sustainability (Basel)*, *12*(18), 7267. doi:10.3390/su12187267

Zhang, L., & Tsung, L. T. (2019). Tibetan bilingual education in Qinghai: Government policy vs family language practice. *International Journal of Bilingual Education and Bilingualism*, *22*(3), 290–302. doi:10.1080/13670050.2018.1503226

Zhang, Y., Liang, X., Wang, Y., & Liang, J. (2018). Natural language understanding for ideological education based on deep neural networks and knowledge graphs. *Jisuanji Yingyong*, *38*(12), 3495–3500.

Zhao, L., Zou, D., Hu, G., Wu, T., Du, E., Liu, G., Xiao, Y., Li, R., Pang, Q., Qiao, Y., Wu, X., Sun, Z., Xing, Z., Sheng, Y., Zhao, Y., Shi, J., Xie, C., Wang, L., Wang, C., & Cheng, G. (2021). A synthesis dataset of permafrost thermal state for the Qinghai-Tibet (Xizang) Plateau, China. *Earth System Science Data*, *13*(8), 4207–4218. doi:10.5194/essd-13-4207-2021

Fae Mai is a expert in the field of natural language modeling and a faculty member at Western Iowa Tech Community College.