RESEARCH ARTICLE

Topic modeling and evolutionary trends of China's language policy: A LDA-ARIMA approach

Tianxin Li[1], **Xigang Ke**[1]*, **Hui Shi**[2]

1 Department of Literature, Shaanxi Normal University, Xi'an, Shaanxi, China, **2** Department of Literature, Nanjing Normal University, Nanjing, Jiangsu, China.

* xigangke@163.com

Abstract

Background

Language policy serves as an essential tool for governments to guide and regulate language development. However, China's current language policy faces challenges like outdated analytical methods, inefficiencies caused by policy misalignment, and the absence of predictive frameworks. This study provides a comprehensive overview of China's language policy by identifying key topics and predicting future trends.

Methods

We employ the Latent Dirichlet Allocation topic model and Autoregressive Integrated Moving Average model systematically analyze and predict the evolution of China's language policy. By gathering a large-scale textual data of 1,420 policy texts from 2001–2023 on official websites, we achieve both topic extraction and evolution prediction.

Results

This study reveals that: (1) Language life, language education, and language resources have high popularity indexes, and language education and language planning exhibit high expected values. (2) The theme intensity of most topics has been a significant upward trend since 2014, with significant fluctuations during T1-T2. (3) From 2001 to 2023, the actual and fitted values show an overall positive trend. In 2024–2028, the predicted value of language resources stabilizes after a brief decline in 2024, while other topics show upward trends.

Conclusions

This study extracts 1,420 policy texts from official websites and outlines the following findings: (1) Language policies focus on maintaining a harmonious linguistic environment, addressing educational inequality, and protecting language resources.

OPEN ACCESS

Citation: Li T, Ke X, Shi H (2025) Topic Modeling and Evolutionary Trends of China's Language Policy: A LDA-ARIMA Approach. PLoS One 20(5): e0324644. https://doi.org/10.1371/journal.pone.0324644

Editor: Li-Pang Chen, National Chengchi University, TAIWAN

Received: September 12, 2024

Accepted: April 27, 2025

Published: May 28, 2025

Data availability statement: "All relevant data are within the paper and its supporting information files. The language policy texts were retrieved from the official Chinese government website (https://www.gov.cn/) by searching for the keyword "语言" (language), as well as from the "国内语情" (Domestic Language Situation) of the General Office of National Language

Commission Research Planning Committee (http://www.ywky.edu.cn/).

Funding: This work was supported by the National Social Science Foundation of China (24VJXG066) , and Social Science Foundation of Shaanxi Province (2023K007). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

(2) Since 2014, most topics have exhibited fluctuating yet sustained growth trend, particularly in language education and research. (3) Except for language resources, the predicted values of the remaining six topics will show a growing trend from 2024 to 2028. Based on these findings, we propose policy recommendations such as strengthening language research, developing a multilingual education system, and optimizing language resource management.

Introduction

Language policy is defined as the laws, regulations, rules, and measures established by human social groups to govern speech communication based on their position and viewpoint on a certain language [1,2]. In China, the language service industry generated 55.45 billion yuan in 2021, representing approximately 16.66% of the global market [3]. The effectiveness of language policy plays a significant role in shaping economic restructuring, driving technological innovation, and maintaining cultural sovereignty [4,5]. However, the rapid digitization of language technologies has exposed a significant gap: China's current language policy frameworks, which primarily rely on retrospective qualitative evaluations [6–8], are insufficiently responsive to the real-time challenges posed by innovation-driven development strategies [9].

Language policies face three key challenges. First, traditional policy analysis methods like expert reviews and manual coding, are increasingly inadequate due to subjective biases, slow feedback, and limited ability to handle large amounts of multilingual data from digital platforms [10,11]. Second, despite the rapid growth of China's language industry, policy misalignment has led to inefficiencies [12], with overinvestment in machine translation while underfunding of minority language preservation [13,14]. Third, as AI-driven language technologies reshape global labor markets, the absence of timely policy adjustments has contributed to workforce disruptions and widening linguistic inequalities in several countries [15,16]. China currently lacks a predictive framework capable of anticipating and mitigating these risks.

Previous research has inadequately addressed these issues. Existing studies on language policy phases remain descriptive, failing to account for the real-time interactions between policy texts and socio-technological outcomes. Although computational methods like topic modeling and time series analysis have been widely applied in fields such as energy policy [17], monetary policy [18], and healthcare policy [19], they have yet to address the complexities of language governance, such as dialect diversity and policy terminology's ideological impact. This gap limits policymakers' ability to assess interventions like AI ethics guidelines or regional language revitalization campaigns.

To bridge this gap, we propose a machine learning framework integrating Latent Dirichlet Allocation (LDA) [20,21] and Autoregressive Integrated Moving Average (ARIMA) models [22]. The main contributions are as follows: (1) Introducing a machine learning method for evaluating language policy that shifts from text analysis to policy effect assessment, offering new methods and perspectives for precise policy

identification. (2) Employing the LDA model to systematically and accurately evaluate key topics and their relationships, enhancing existing literature. (3) Applying ARIMA model to enable real-time monitoring and feedback on language policy implementation, aiding policymakers in making prompt adjustments and enhancements to ensure policy efficacy and flexibility.

Literature review

Progress in language policy and text mining

Language policy plays a crucial role in managing national society, resolving political conflicts and centralizing political resources [23–25]. In China, early language policy efforts focused on macro-level planning [26], including implementing comprehensive language and writing reforms, such as the organization and simplification of Chinese characters and the widespread promotion of Mandarin in 1949 [27]. In the 21st century, language policy has gradually shifted from a macro to a micro focus, addressing linguistic diversity and social needs [28]. Most existing studies on language policy concentrate on policy implementation and evaluation [29–31]. Research on policy implementation explore the evolution of language policy [32], social network analysis [33], discourse analysis [34], and cross-national comparisons [35]. Policy evaluation studies, on the other hand, focus on language return rates [36], the socio-economic effects of language policy [37], cost-benefit analysis [38], and future prospects [39].

Despite these advancements, much research remains largely qualitative and descriptive, relying on case studies and expert analysis. Moreover, large-scale, data-driven methods are underutilized for dynamic policy impact assessment. Text mining, a computational technique for extracting meaningful patterns from unstructured texts, has gained increasing attention in policy research [40]. Scholars have shown great interest in large-scale policy documents, including government policy texts, legal records, policy news and media data [41–43]. Common text mining techniques such as text clustering, text classification, and feature extraction have been widely used and validated [44–46]. Policy text mining finds applications in text information retrieval, sentiment analysis, theme evolution analysis, and performance evaluation [47–49]. Natural language processing (NLP) methods based on big data can transform policy texts into structured data, facilitating the construction of specialized objects such as semantics or sentiment [50,51].

However, the application of text mining in language policy research remains underdeveloped. While its successful implementation in energy regulation, economic policy, and healthcare policy [52–54], language policy presents unique challenges, including dialect diversity, ideological nuances, and regional variations. Some studies focus on short-term policies, limiting their ability to predict shifts and assess long-term impacts [55]. Furthermore, many studies rely on qualitative methods [56], which may lead to a lack of objectivity. Therefore, there is a need to incorporate quantitative methods for more precise analysis. Utilizing NLP techniques based on big data can offer a more systematic, scalable, and objective evaluation of policy effectiveness [57].

Topic modeling and evolutionary trends

Topic modeling extracts key content from texts, uncovers hidden policy insights, and identifies topic relationships [58,59]. This process aids in enabling individuals to promptly comprehend the trends of language policy development, providing the basis for in-depth analysis of language issues and strengthening the role of policy on social and economic development.

At present, there are a large number of academic research, primarily focusing on topic modeling and visualization [60–62]. Scholars have concentrated on the current state of topic modeling, such as the work of Blei, an American academic, who utilized the LDA topic model to reveal hidden subject information within extensive texts or corpora [20]. Furthermore, some researchers have analyzed the temporal dynamics in the socio-political landscapes during presidential elections, analyzing political texts to explore the multi-faceted nature of public policy [63]. Other scholars have expanded

on evolutionary analysis by integrating the weighted Jacobian matrix with LDA analysis to illustrate the trend of topic change [64].

As research has progressed, several studies have sought to better capture topic evolution over time, though challenges remain. In response, scholars have made notable attempts to address this gap. For instance, after constructing theme models, some scholars have utilized sentiment classification models to categorize the derived topics [65]. Several studies employed the ARIMA model to forecast the derived topics [66]. Others have combined time series analysis with support vector machine model, generating time series data through LDA topic modeling and using support vector machine model for trend analysis [67,68]. Furthermore, combining topic modeling with evolutionary trends presents challenges. One issue is that topics evolution is often nonlinear, making it difficult for traditional time series models to capture their shifts. In addition, the integration of qualitative thematic content with quantitative forecasting introduces complexity, highlighting the need for more advanced approaches that consider both semantic features and temporal dynamics.

In summary, most existing studies focus on the visualization analysis of topic modeling. While some studies have started to explore evolutionary trends, the predominant approach involves constructing time series models and interpreting these trends manually. To date, few research has combined topic evolution and time series analysis on language policy. Given the strengths of LDA model, such as its strong dimensionality reduction capabilities [69], robust foundation in probability theory [70], and scalability [71], along with the ARIMA model's accurate short-term predictive capabilities for analyzing time-dependent data, this paper aims to bridge this gap. Specifically, we collect China's language policy texts from official websites from 2001 to 2023. The research employs the LDA theme model to identify key topics and constructs an ARIMA model to predict the theme intensity of language policy from 2024 to 2028. This integrated approach synthesizes internal semantic features and external theme intensity to analyze theme evolution trends and provide predictive assessments.

Methods

Dataset

The study obtained language policy texts from the official website of the Chinese government (https://www.gov.cn/) using the keyword "语言" (language) and from the "国内语情" (Domestic Language Situation) of the General Office of the National Language Commission Research Planning Committee (http://www.ywky.edu.cn/) from January 2001 to October 2023. Both sources are authoritative Chinese government websites. After removing incomplete, redundant, or irrelevant texts, 1,420 policy texts were selected as the main research sample, ensuring the dataset's integrity and quality.

Data preprocessing

Thematic analysis typically requires data preprocessing to refine the data, which includes three key tasks:

(1) Data cleaning. We remove extraneous content like HTML tags, special characters, noise, null values and duplicate entries from the policy text to ensure the data's consistency and suitability for topic modeling. In addition, we convert relevant English terms to lowercase, while keeping key numerical values for context.

(2) Text Segmentation. Because Chinese lacks explicit word boundaries, we employ Python's Jieba tokenizer for segmentation, ensuring accurate identification of multi-character policy terms.

(3) Punctuation and stopword removal. To enhance the quality of text mining, we remove punctuation marks, such as period (“.”), comma (“,”), question mark (“?”), exclamation point (“!”), and special characters including ampersand (“&”) and slash (“/”). In addition, we eliminate stopwords—such as “和”, “是” and “的”—that do not contribute to semantic meaning. The stopword list is sourced from a public website (https://countwordsfree.com/stopwords).

Utilizing latent dirichlet allocation for topic identification

This study employs LDA topic modeling for thematic analysis of language policy texts. LDA, an unsupervised machine learning method, uses generative probabilistic techniques to deduce word distributions that characterize various topics [72]. This approach allows for the examination of policy texts with diverse topics and complex lexical features, converting textual information into a digital format. The fundamental equation is denoted as (1):

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)\, p(w_n|z_n, \beta) \tag{1}$$

Where θ represents the topic distribution for a document, while **z** and **w** denote the sequence of topic assignments and the sequence of words in the document. The hyperparameters α and β govern the Dirichlet prior distributions, $p(\theta|\alpha)$ models the probability of the topic distribution conditioned on the Dirichlet prior α. For each word in the document, $p(z_n|\theta)$ represents the probability of assigning topic z_n to word w_n, and $p(w_n|z_n, \beta)$ denotes the probability of generating word w_n given its assigned topic z_n, and the word distribution parameter β. The product notation ∏ accounts for all words n in the document.

We preprocess the data to create a corpus for LDA analysis, following these steps: (1) We vectorize policy texts using the Bag-of-Words (BOW) model and apply Term Frequency-Inverse Document Frequency (TF-IDF) weighting to refine word importance. (2) We train an initial LDA topic model while determining the optimal number of topics by minimizing perplexity and maximizing coherence, ensuring topic quality and interpretability [73,74]. (3) Using the optimal number of topics, we train the final LDA model with the selected hyperparameters, producing topic-word and document-topic distributions. (4) We calculate the theme intensity over time for each topic to analyze the evolving trends in language policy.

Evolutionary trends based on autoregressive integrated moving average model

ARIMA, a widely used statistical method for time series forecasting, which involves curve fitting and parameter estimation to develop a mathematical model for future predictions. It is applied in various fields, including economics, finance, hydrological forecasting, and aerospace [75–78]. We apply the ARIMA model to forecast theme intensity, represented as ARIMA (p, d, q), where p represents the autoregressive term, d signifies the difference order, and q denotes the number of moving average terms. The calculation is detailed in Equation (2).

$$\Delta^d y_t = \alpha \sum_{i=1}^{p} \gamma_i \Delta^d y_{t-i} + \beta \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t + \mu \tag{2}$$

Where $\Delta^d y_t$ is the d-th order differenced value of the original time series y_t, γ_i is the autocorrelation coefficient, θ_i is the moving average coefficient, ε_t is the error, μ is the constant term, and α and β are the coefficients. The process of using the ARIMA model to analyze theme intensity evolution involves three key steps:

(1) Stationarity test. We assess stationarity using the Augmented Dickey-Fuller (ADF) test, where the null hypothesis assumes a unit root (non-stationarity). If the ADF statistic is higher than the 5% critical value, the series is non-stationary, and first-order differencing is applied. This process is repeated until stationarity is confirmed.

(2) Parameter estimation. To estimate model parameters, we first examine the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), though subjective factors may influence interpretation. Following the Box-Jenkins methodology, we complement this approach with the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) [79,80]. To balance forecasting accuracy and model complexity, we explore the autoregressive (AR)

and moving average (MA) orders over the ranges $0 \le p \le 2$, and $0 \le q \le 2$. The optimal parameter is selected based on $\frac{(AIC+BIC)}{2}$ minimum [81].

(3) Model validation and prediction. We validate model adequacy by applying the Ljung-Box test to residuals [82], where $p > 0.05$ indicates white noise. Predictive performance is assessed using out-of-sample rolling forecasts, measured by Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The final ARIMA(p,d,q) model is then used to forecast trends for the next five years. The entire process is shown in Fig 1.

Results

Determining the optimal number of topics

To determine the optimal number of topics, researchers have explored several quantitative and qualitative methods such as perplexity, coherence [83,84]. This study integrated perplexity and coherence to identify the optimal number of topics. Fig 2 demonstrated that LDA exhibited lower perplexity and higher coherence when $K = 7$, suggesting that the optimal number of topics was 7.

Policy text keywords analysis

The word cloud in Fig 3 illustrated high-frequency keywords central to language policy: language, common language, research, country, development, culture, education, construction, and service. These keywords collectively outlined the core aspects of language policy, encompassing research, education, and services. In research, language policies addressed language resources, cognition, and life. In education, policies focused on language proficiency, schools,

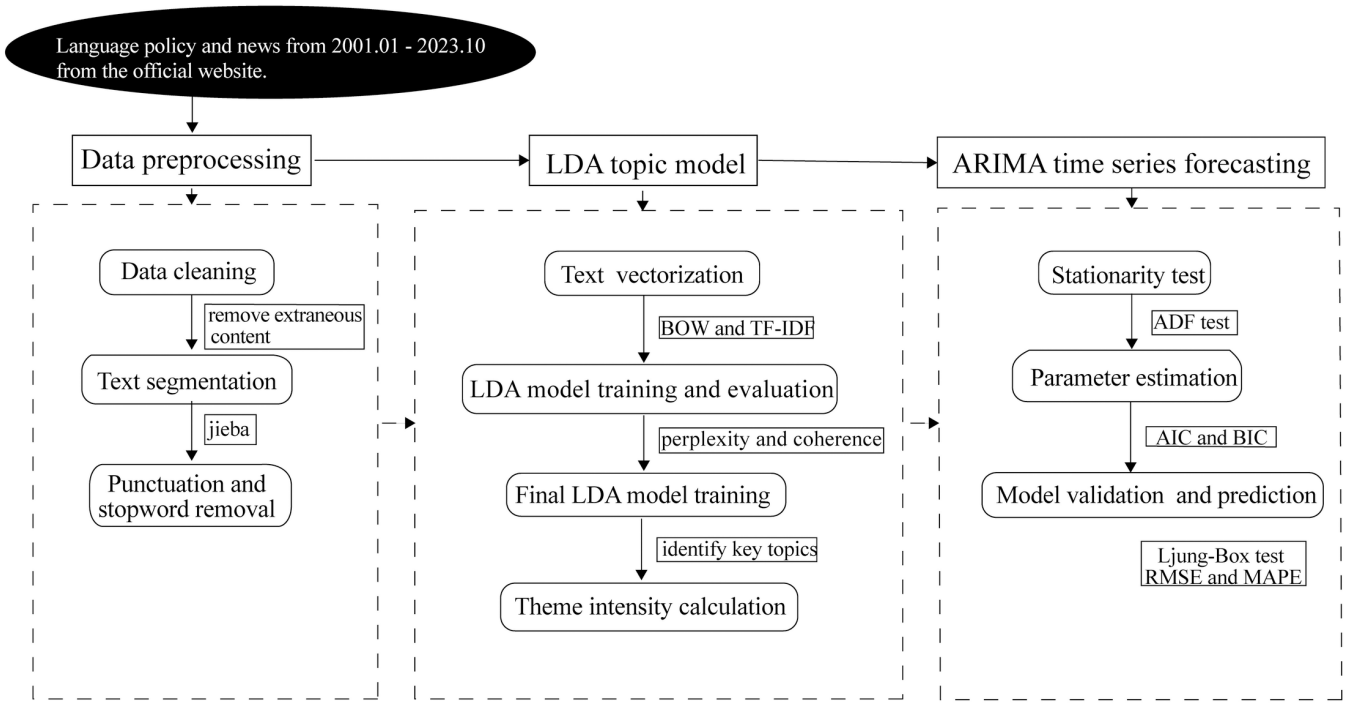

Fig 1. The flowchart of the proposed method.

https://doi.org/10.1371/journal.pone.0324644.g001

students, common language, and dialects. Regarding services, they emphasized the integration of translation services for English and Chinese, as well as AI technologies.

Topic modeling results

Topic identification and popularity analysis. To identify research directions and hotspots, Table 1 displayed the parameters related to the probability distribution, including popularity index, expected value, standard deviation and dispersion coefficient. We had the following findings: (1) Topic 1 had the highest popularity index of 0.318, followed

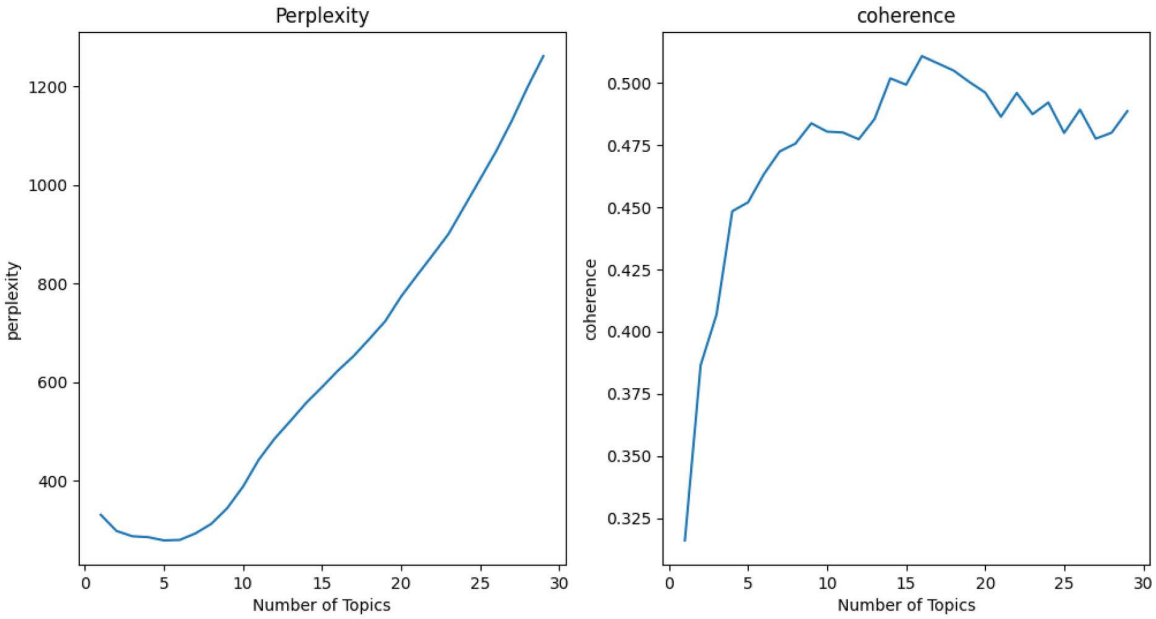

Fig 2. Perplexity and coherence of LDA models with different values of K[a]. [a] The figure examines the range of topics K from 0 to 30, utilizes a step size of 5 for LDA topic extraction, and evaluates perplexity and coherence on the test dataset.

https://doi.org/10.1371/journal.pone.0324644.g002

Fig 3. The high-frequency word cloud of China's language policy.

https://doi.org/10.1371/journal.pone.0324644.g003

by Topic 2 and 3. However, Topic 6 and 7 had the lowest popularity index of 0.045, indicating their marginal status. (2) Most topics exhibited relatively high expected values, except for Topic 7. Notably, Topic 2 and 6 had the highest expected values of 0.154, warranting further exploration of their unique attributes. (3) Topics 2, 3 and 6 had higher dispersion coefficients, indicating that language education, language resources and language planning exhibited broader distributions and potentially greater diversity.

Thematic content analysis. In Table 2, the top 15 key terms for the seven topics were systematically ranked to further analyze their relevance.

Topic 1: Language life. Topic 1 featured terms such as engineering, network, norms, diversity, and media, highlighting the important role of digitization in promoting harmonious language development. Engineering and network exhibited the highest relevance of 0.007, associated with the language resources protection project initiated in 2015.

Topic 2: Language education. Topic 2 encompassed terms such as textbook, construction, governance, policy, and disciplines, emphasizing the collaborative efforts of government, education, and media entities. Textbook exhibited the highest relevance of 0.007, while construction and governance followed with a relevance of 0.006, highlighting the significance of language education, governance support and technological advances.

Topic 3: Language resources. Topic 3 reflected research priorities in the development and preservation of language resources. The relevance for Chinese and network were both 0.008, indicating Chinese language resources play an important role in the digital sphere.

Topic 4: Language protection. Topic 4 contained terms such as protection, language, communication, theory, and native Language, with protection garnering the highest relevance of 0.007. This topic delved into the objectives, current status, and significance of language protection, emphasizing the necessity of effective mechanisms to support the development and implementation of language protection initiatives.

Topic 5: Language research. Topic 5 covered terms like strategy, civilization, linguistic, monitoring, college, and academic, with strategy exhibiting the highest relevance of 0.007. This topic concentrated on different aspects of language research, including methodology, content, and institutional frameworks.

Topic 6: Language planning. Topic 6 included terms such as language policy, think tank, planning, capacity, and era, with language policy having the highest relevance of 0.009. The topic centered on the issues, content, progress, and organizational aspects of language planning, ranging from language policy to public linguistic life.

Topic 7: International communication. Topic 7 involved terms like Belt and Road, university, intelligence, seminar, and minority, with Belt and Road having the highest relevance of 0.010. This topic highlighted the significance of language in international communication, particularly emphasizing the Belt and Road initiative's role in this context.

Overall, several key items were consistent with the word cloud, with particular attention to diversity, governance, grammar, Belt and Road, and strategies in the thematic analysis. This reinforced the emphasis on linguistic diversity, language planning and language education as priority areas.

Table 1. Statistical parameters of China's language policy.

	Topics	Popularity index	Expected value	Standard deviation	Dispersion coefficient	Rank
1	Language life	0.318	0.153	0.002	0.015	1
2	Language education	0.227	0.154	0.009	0.056	2
3	Language resources	0.182	0.152	0.004	0.029	3
4	Language protection	0.091	0.151	0.001	0.007	4
5	Language research	0.091	0.153	0.001	0.001	5
6	Language planning	0.045	0.154	0.005	0.030	6
7	International communication	0.045	0.001	0.001	0.001	7

https://doi.org/10.1371/journal.pone.0324644.t001

Theme intensity evolution analysis

Fig 4 illustrated the trend of theme intensity for each topic, highlighting shifts in China's language policy focus over different time periods. There were the following results: In general, the theme intensity of most topics has exhibited an upward trend since 2014, with significant fluctuations during T1-T2. Firstly, Topic 1 has shown an increasing trend since 2015, experiencing a decline to 0.149 in 2021 before rising again. This trend pattern corresponded with the release of the *13th Five-Year Plan for the Development of the National Language and Literature Program* and the *Implementation Plan for the Universalization of the State Common Language and Literature Project*.

Table 2. Terms and their relevance for theme analysis[a].

	Topics	Term	Relevance	Term	Relevance	Term	Relevance
1	Language life	engineering	0.007	language life	0.006	languages	0.005
		network	0.007	poverty alleviation	0.005	grammar	0.005
		norms	0.006	methods	0.005	standard	0.004
		diversity	0.006	language resources	0.005	language protection	0.004
		media	0.006	advancement	0.005	professor	0.004
2	Language education	textbook	0.007	platform	0.005	school	0.005
		construction	0.006	university	0.005	French	0.004
		governance	0.006	education	0.005	consciousness	0.004
		policy	0.005	important	0.005	children	0.004
		disciplines	0.005	artificial intelligence	0.005	Chinese	0.004
3	Language resources	Chinese	0.008	language ability	0.006	center	0.005
		network	0.008	dialect	0.006	Chinese character	0.005
		center	0.007	director	0.006	collaborative	0.005
		language resources	0.007	demand	0.006	condition	0.005
		Mandarin	0.006	language and information division	0.005	vocabulary	0.005
4	Language protection	protection	0.007	native language	0.006	global	0.005
		language	0.006	person	0.006	terms	0.005
		communication	0.006	head	0.005	survey	0.005
		linguistic	0.006	institute	0.005	function	0.005
		theory	0.006	research	0.005	discussion	0.005
5	Language research	strategy	0.007	academic	0.005	English	0.005
		civilization	0.006	translation	0.005	center	0.005
		linguistic	0.006	development	0.005	cognitive	0.004
		monitoring	0.006	research	0.005	way	0.004
		college	0.005	subject	0.005	theory	0.004
6	Language planning	language policy	0.009	dialect	0.006	national languages	0.005
		think tank	0.008	linguistics	0.005	phenomenon	0.005
		planning	0.007	talent	0.005	analysis	0.005
		capacity	0.007	Mandarin	0.005	English	0.004
		era	0.006	community	0.005	forum	0.004
7	International communication	Belt and Road	0.010	national language commission	0.005	culture	0.005
		university	0.007	scientific research	0.005	human	0.005
		intelligence	0.007	informatization	0.005	America	0.005
		seminar	0.007	system	0.005	nation	0.005
		minority	0.006	scholar	0.005	world	0.005

[a] The list of terms was derived from the LDA results using Gensim (K = 7) to automatically extract semantic topics.

https://doi.org/10.1371/journal.pone.0324644.t002

Secondly, Topic 2 has shown continuous growth since 2014, reaching 0.154 in 2023, highlighting China's emphasis on language education. The government launched the *National Universal Language and Script Popularization Project* in 2017, formulated the *Opinions on Comprehensively Strengthening Language and Literature Work in the New Era* in 2020 and proposed the *full implementation of education and teaching in the national common language and script* in 2022. These measures have significantly shaped language policy, enhancing both the accessibility and quality of language education. Topic 3, after experiencing a brief decline in 2015, peaked in 2019 and 2021 with a maximum value of 0.152. The first phase of the China Language Resources Protection Project concluded in 2019, and the launch of its second phase in 2021 reinforced efforts in language data mining and informed decision-making.

Thirdly, Topic 4 exhibited fluctuations before peaking at 0.151 in 2020, highlighting policymakers' focus on language protection. Topic 5 fluctuated and grew, peaking at 0.151 in 2017 and 0.152 in 2019. Topic 6 showed a fluctuating upward trend, stabilizing at 0.151 after 2020. Finally, Topic 7 experienced fluctuating growth, peaking at 0.151 in 2020 after a brief decline in 2017. In 2020, the Ministry of Education established the *Center for Sino-foreign Language Exchange and Cooperation* and the *China International Chinese Language Education Foundation*. These initiatives have promoted cultural exchanges and enhanced international understanding.

Prediction of topic evolutionary trends

Time series difference processing and test. The ADF test is widely employed as a prevalent unit root test for evaluating the stationarity of time series data [85]. Table 3 indicated that the p-value for all topics, except Topic 5, exceeded 0.05, indicating non-stationarity. However, the p-value dropped below 0.05 after differencing, suggesting that stationarity was achieved. Topic 5 passed the ADF test, indicating stationarity at the 0.1% significance level, while the remaining topics exhibited stationarity in their time series after differencing.

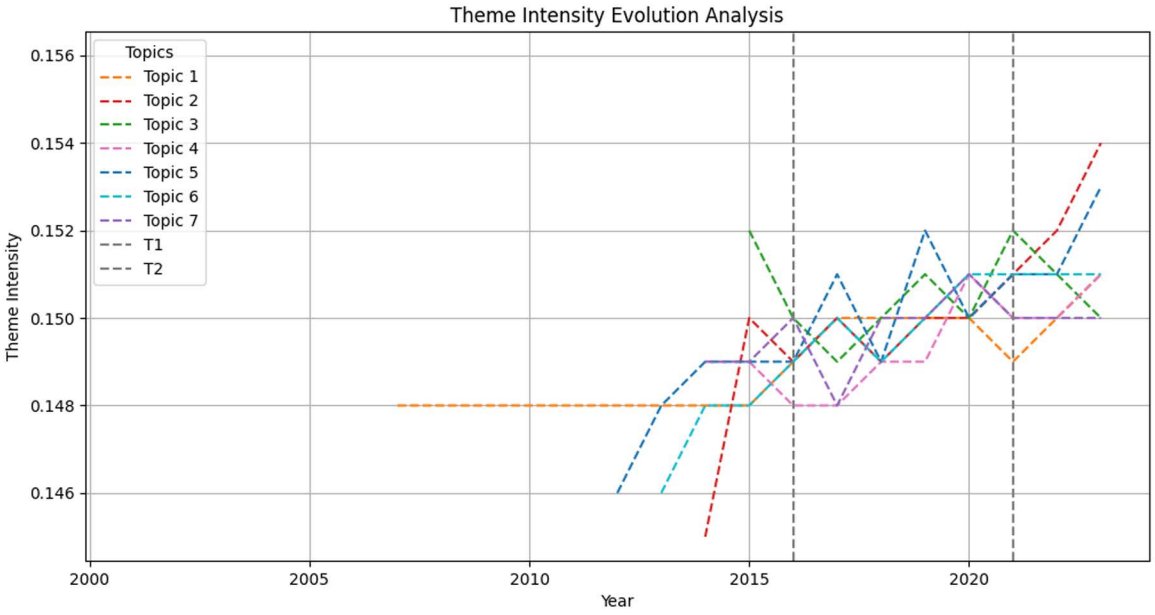

Fig 4. Theme intensity evolution analysis[a]. [a] T1 denotes the year 2016, while T2 signifies 2021. These specific time points were selected due to the first phase of the language resources protection project in China in 2016. Moreover, 2021 marks the subsequent phase in the development of the language resources protection project.

https://doi.org/10.1371/journal.pone.0324644.g004

Parameter estimation and model validation. We performed the Ljung-Box test on the residual series to assess the robustness of the fitting and predictive abilities for each topic in Table 4. Firstly, we selected the parameter with the $\frac{(AIC+BIC)}{2}$ minimum to establish optimal models for Topic 1–7. Secondly, the Ljung-Box test results for all topics exhibited $p > 0.1$, indicating that there was no significant autocorrelation. Specifically, Topic 1 addressed trend effects through second-order differencing. The optimal model, ARIMA(2,2,0), was identified with parameters $p = 2$ and $q = 0$, achieving the $\frac{(AIC+BIC)}{2}$ minimum (AIC = -133.565, BIC = -131.305). The Ljung-Box Q(6) test yielded a p-value of 0.314, confirming the absence of significant autocorrelation and validating the model's suitability.

Results of topic evolutionary trends prediction. To understand the dynamics of China's language policy, we analyzed the theme intensity within the raw time-series data from 2001 to 2023 employing an ARIMA model, then we predicted the trend of seven topics from 2024 to 2028, as shown in Fig 5.

Regarding the actual values, Topics 1, 2, 4, 5, 6 and 7 exhibited an increasing trend from 2001 to 2023. Topic 2 exhibited the highest increase, rising from 0.145 to 0.154, with an increase of 0.009. While Topic 3 showed a fluctuating trend,

Table 3. Time series ADF test results[a].

Topics	Difference	t-statistic	Test critical values			p-value
			1% level	5% level	10% level	
Topic 1	d = 0	-1.943	-4.380	-3.600	-3.240	0.632
	d = 2	-4.609	-4.380	-3.600	-3.240	0.001***
Topic 2	d = 0	-4.277	-4.380	-3.600	-3.240	0.003*
	d = 1	-8.203	-4.380	-3.600	-3.240	0.001***
Topic 3	d = 0	-3.093	-4.380	-3.600	-3.240	0.108
	d = 2	-4.486	-4.380	-3.600	-3.240	0.002**
Topic 4	d = 0	-3.279	-4.380	-3.600	-3.240	0.070
	d = 2	-3.490	-4.380	-3.600	-3.240	0.041*
Topic 5	d = 0	-5.632	-4.380	-3.600	-3.240	0.001***
Topic 6	d = 0	-2.575	-4.380	-3.600	-3.240	0.292
	d = 3	-3.451	-4.380	-3.600	-3.240	0.045*
Topic 7	d = 0	-3.204	-4.380	-3.600	-3.240	0.084
	d = 1	-4.563	-4.380	-3.600	-3.240	0.001***

[a] *p < 0.05, **p < 0.01, ***p < 0.001.

https://doi.org/10.1371/journal.pone.0324644.t003

Table 4. ARIMA model parameter estimation and Ljung-Box test results.

Topics	ARIMA(p,d,q)	p-value	AIC[a]	BIC[a]	Ljung-Box Q(6)[b]
Topic 1	ARIMA(2,2,0)	0.001***	-133.565	-131.305	0.314
Topic 2	ARIMA(1,1,0)	0.001***	-72.557	-72.318	0.501
Topic 3	ARIMA(1,1,0)	0.050*	-49.500	-50.124	0.728
Topic 4	ARIMA(2,1,0)	0.004**	-52.253	-53.086	0.581
Topic 5	ARIMA(1,0,0)	0.006**	-118.921	-117.226	0.929
Topic 6	ARIMA(1,1,0)	0.030*	-37.918	-39.090	0.513
Topic 7	ARIMA(1,1,0)	0.010**	-87.031	-86.439	0.151

[a] The model that exhibited the most optimal fit and featured all model parameters that were statistically significant was chosen by employing the criteria of $\frac{(AIC+BIC)}{2}$ minimum.

[b] The Ljung-Box test with $p > 0.05$ did not provide sufficient evidence to reject the null hypothesis that there is no autocorrelation.

https://doi.org/10.1371/journal.pone.0324644.t004

peaking at 0.152 in 2022 before gradually decreasing. In terms of the fitted values, Topics 1, 2, 3, 4, 6 and 7 exhibited an overall upward fluctuation from 2001 to 2023. Topic 4 peaked at 0.151 in 2022 before stabilizing, while Topic 5 remained relatively stable at 0.150. Fig 5 illustrated the similarity between the fitted series and the actual series, verifying the overall effectiveness of the model in capturing the time series of theme intensity.

In terms of predicted values, Topics 1, 2, 4, 5, 6 and 7 exhibited positive trends between 2024 and 2028, with Topic 2 showing the fastest growth. Meanwhile, Topic 3 was expected to gradually stabilize following a brief decline from its peak of 0.152 in 2023. In summary, the predicted values for all seven topics fell within the 95% confidence interval, indicating the high accuracy of the model in forecasting the data.

Evaluation of autoregressive integrated moving average model efficacy. To assess the efficacy of the ARIMA model, RMSE and MAPE were used as key metrics for evaluating model performance [86,87], as detailed in Table 5. The findings indicated that the RMSE values approached zero across all seven topics, while the MAPE values consistently remained below 5%. These results highlighted the model's strong predictive accuracy, affirming the method's effectiveness.

Discussion

Language policy serves as a crucial instrument for governments to regulate and guide language development, thereby attracting significant academic interest [88–90]. This study employs the LDA unsupervised machine learning method and the ARIMA model with high predictive accuracy to analyze 1,420 China's language policy texts. The analysis identifies seven topics and forecasts their evolutionary trends over the next five years, aiming to enhance the effectiveness of language policy.

LDA is an effective large-scale thematic model for textual analysis used to assess the topics and intensities embedded in policy texts [91]. Topic 1 has the highest prevalence index of 0.318 and a high expectation value of 0.153, indicating significant social concern about language use and development. The thematic content analysis emphasizes the importance of language texts, knowledge, and emerging technologies in shaping language life. The trend of theme intensity evolution has shown a continuous increase since 2015, with a brief decline in 2021, followed by a renewed rise. In 2022, China achieved an 80.72% penetration rate of Mandarin, showcasing advancements in promoting the national language [92]. As China works towards standardizing its language, its linguistic values are also diversifying, resulting in a rich and colorful language life where various languages and dialects coexist, reflecting the characteristics of the times.

Topic 2 has the second-highest popularity index of 0.227 and the highest expectation value of 0.154, highlighting its significance in language policy. The thematic content analysis reveals the significant role that textbooks and discipline building in enhancing the quality of language education. Scientific language policy and management serve as powerful tools for promoting educational reform and enhancing cultural literacy [93]. The theme intensity has shown continuous growth since 2014, reaching 0.154 in 2023, indicating China's strong emphasis on language education. In 2020, the proportion of the literate population using standardized Chinese characters exceeded 95.00%, and the illiteracy rate dropped to 2.67% [94]. Despite significant progress, challenges persist, such as the limited variety of teaching materials and the insufficient quality of teaching staff [95,96]. Therefore, it is imperative to adjust and enhance language education policies, implement reforms to optimize academic disciplines development.

The high popularity index of 0.182 for Topic 3 underscores the importance of the Chinese on digital platforms, as well as the crucial role of language proficiency, dialects, and Mandarin in fostering sustainable language resource development. The theme intensity of language resources fluctuated in stages, reflecting gradual adjustments in China's language resource protection initiatives. At present, China has established the world's largest language resource database, encompassing over 120 languages and dialects [97]. While the development of language resources can yield economic benefits, it also faces challenges, such as balancing the relationship between common languages and dialects and ensuring the protection of vulnerable language groups.

Fig 5. Results of the trend prediction of China's language policy[a]. [a] The UCL and LCL confidence levels are obtained by extrapolating the sample data, where UCL denotes the high limit of the confidence interval and LCL denotes the low limit of the confidence interval.

https://doi.org/10.1371/journal.pone.0324644.g005

Table 5. ARIMA model predictive assessment.

Predictive model	Topic 1 ARIMA(2,2,0)	Topic 2 ARIMA(1,1,0)	Topic 3 ARIMA(1,1,0)	Topic 4 ARIMA(2,1,0)	Topic 5 ARIMA(1,0,0)	Topic 6 ARIMA(1,1,0)	Topic 7 ARIMA(1,1,0)
RMSE	0.003	0.004	0.004	0.002	0.005	0.002	0.003
MAPE(%)	1.667	2.686	2.630	1.418	3.104	1.169	2.006

https://doi.org/10.1371/journal.pone.0324644.t005

Topic 4 emphasizes language preservation through a comprehensive policy strategy aimed at sustaining linguistic vitality. The intensity peaked at 0.151 in 2020. In recent years, the government has protected languages and scripts by enacting laws, setting up dedicated agencies, and executing projects to protect language resources [98,99]. Despite significant progress in language preservation, challenges persist, including the great variety of languages and the rapid rate of endangerment. Exploring innovative technologies and methods is crucial to address these issues.

Topic 5, with an expectation value of 0.153, focuses on the methods and content of language research, highlighting the importance of language strategies and monitoring to understand the state of languages and support language policy development. The trend of theme intensity showed fluctuating growth, indicating sustained interest in language research. However, current language research tends to favor analysis, neglecting the study of real discourse and language function [100,101]. Effectively addressing the complexities of language research demands a multidisciplinary approach.

Topic 6, with the highest expectation value of 0.154, explores the benefits of intelligent language construction through AI and big data. The theme intensity has stabilized since 2020. The Chinese government has strengthened the management system for language and writing recently. More than 80 laws and regulations, such as the *Law on the State Common Language and Writing System* and *administrative rules concerning the Putonghua Shuiping Ceshi*, have been established [98,102]. These efforts in language planning have provided significant support and guarantees for economic, social, and linguistic development.

Topic 7 highlights the importance of language in international exchange, particularly the strategic role of the Belt and Road Initiative. The theme intensity has fluctuated and grown, peaking at 0.151 in 2020. By 2019, the number of Chinese language learners worldwide had surpassed 100 million, with more than 500 Confucius Institutes established globally [103]. This underscores the need to integrate economic growth with strengthened bilateral and multilateral language and cultural exchanges to foster international understanding and cooperation.

The ARIMA model was used to predict the trends of seven topics over the next five years. Fig 5 shows that, with the exception of language resources (Topic 3), the predicted values for the other six topics will exhibit an upward trend from 2024 to 2028. Notably, the predicted value for language education (Topic 2) is expected to show the most significant growth. In contrast, Topic 3 will stabilize after a brief decline, while the predicted value for language research (Topic 5) will remain steady at 0.150. Language education has consistently attracted considerable attention, and projections suggest this trend will persist. This underscores the significance of language education in language policy, particularly the role of foreign language education in promoting language policy implementation and international communication. Language resources peaked in 2021, attributed to the initial success of the language protection project, which heightened government and societal attention to the development and utilization of language resources. However, model projections indicate that language resources will gradually level off after a brief decline in 2023. This trend is attributed to rapid economic growth and urbanization, which pose risks of language endangerment or disappearance, especially in economically developed cities and regions. The relatively stable predicted value for language research indicates sustained interest in this area. Language research primarily focuses on topics such as language governance, language translation, and research methodologies, reflecting the capacity of language policy to provide technical and strategic support.

Based on the obtained discussion, this study proposes the following optimization paths.

Firstly, establish a collaborative language education system involving the government, schools, and society. The government should integrate language education into its top-level design and provide public services. Schools must advance

inclusive language education, bridging urban-rural gaps. Society should organize regular language education activities using existing resources. This cooperation is essential for developing balanced and sustainable language policies.

Secondly, robust language research is crucial for fostering a harmonious linguistic life and refining language planning. Advancements in language research can support linguistic diversity and enhance language governance through initiatives such as promoting common language, surveying dialect and protecting endangered languages.

Thirdly, strengthening regional connectivity is essential for optimizing language resource management. Given China's regional economic variations and different urban backgrounds, it is imperative to develop region-specific language industries that fully utilize local language resources.

Conclusions and limitations

This study investigated 1,420 policy texts from official government websites, extracting seven topics using LDA and ARIMA models. The study drew the following conclusions: (1) Language life, language education, and language resources hold a prominent place in China's language policy, indicating the country's commitment to building a harmonious linguistic environment, addressing educational inequality, and protecting language resources. (2) Since 2014, most topics have shown an upward trend, especially language education and language research, reflecting significant efforts by the Chinese government to achieve economic and social benefits. (3) Prediction trends indicate that, except for language resources, the predicted values of the other six topics will show a positive trend from 2024 to 2028, with language education demonstrating particularly strong growth. Based on these findings, we propose policy recommendations such as deepening language research, developing a multilingual education system, and optimizing language resource management.

However, there are still some limitations: (1) Topic modeling limitations. Thematic extraction results are presented in an abstract or data-oriented manner, requiring integration with relevant contextual backgrounds for full comprehension. However, this focus on identifying innovative research points may result in biased conclusions, potentially neglecting broader research fields. (2) ARIMA model limitations. The ARIMA model, despite its simplicity and interpretability, displays sensitivity to data outliers and does not account for noise, the characterization of which remains unclear. (3) Data limitations. The primary dataset comprised official policies and news, which could introduce validation bias. Addressing such bias falls beyond the scope of this study.

Acknowledgments

T.L. acknowledges the support of the China Scholarship Council program (project ID: 202406870061).

Author contributions

Conceptualization: Tianxin Li, Hui Shi.

Data curation: Tianxin Li, Xigang Ke.

Formal analysis: Tianxin Li, Hui Shi.

Project administration: Tianxin Li, Xigang Ke, Hui Shi.

Writing – original draft: Tianxin Li, Xigang Ke.

Writing – review & editing: Tianxin Li.

References

1. Spolsky B. Language policy. Cambridge University Press. 2004.
2. Ricento T. Theoretical perspectives in language policy: an overview. An introduction to language policy: theory and method. 2006:3–9.

3. Xu R. Breakthroughs and challenges of machine translation in the era of artificial intelligence. In: 2023 8th International Conference on Information Systems Engineering (ICISE). IEEE. 2023. 520–3. https://doi.org/10.1109/icise60366.2023.00116
4. Liddicoat AJ. Language-in-education policies: The discursive construction of intercultural relations. Multilingual Matters. 2013.
5. Wiley TG. Language planning and policy. Sociolinguistics and language teaching. 1996:103–47.
6. Luo H, Meng Y, Lei Y. China's language services as an emerging industry. Babel. 2018;64(3):370–81. https://doi.org/10.1075/babel.00043.luo
7. Andrews E. Language planning in the post-communist era: The struggles for language control in the new order in Eastern Europe, Eurasia and China. Springer. 2018.
8. Li X, Shen Q. Individual agency in language-in-education policy: a story of Chinese heritage language schools in multilingual Brussels. Curr Issues Lang Plan. 2023;25(2):137–56. https://doi.org/10.1080/14664208.2023.2259154
9. Zhang C, Guan J. How policies emerge and interact with each other? A bibliometric analysis of policies in China. Sci Public Policy. 2022;49(3):441–59. https://doi.org/10.1093/scipol/scab091
10. Ye J, Jing X, Li J. Sentiment analysis using modified LDA. In: Signal and Information Processing, Networking and Computers: Proceedings of the 3rd International Conference on Signal and Information Processing, Networking and Computers (ICSINC). Springer Singapore. 205–12.
11. Feng A, Adamson B. Language policies and sociolinguistic domains in the context of minority groups in China. J Multiling Multicult Dev. 2017;39(2):169–80. https://doi.org/10.1080/01434632.2017.1340478
12. Barr DM. Who's afraid of China?: the challenge of Chinese soft power. Bloomsbury Publishing. 2011.
13. Hutchins J. Current commercial machine translation systems and computer-based translation tools: system types and their uses. Int J Transl. 2005;17(1–2):5–38.
14. Wang Y, Ye H, Schapper A. Multiculturalism and multilingual education for minority ethnic groups in China: Examples of Southwest China and Xinjiang Uygur regions and the goal of educational equality. Learning from Difference: Comparative Accounts of Multicultural Education. 2016: 51–68. https://doi.org/10.1007/978-3-319-26880-4_4
15. Calo R. Artificial intelligence policy: a primer and roadmap. UCDL Rev. 51:399.
16. Keppo I, Rao S. International climate regimes: Effects of delayed participation. Technol Forecast Soc Change. 2007;74(7):962–79. https://doi.org/10.1016/j.techfore.2006.05.025
17. Benites-Lazaro LL, Giatti L, Giarolla A. Topic modeling method for analyzing social actor discourses on climate change, energy and food security. Energy Res Soc Sci. 2018;45:318–30. https://doi.org/10.1016/j.erss.2018.07.031
18. Luangaram P, Wongwachara W. More than words: a textual analysis of monetary policy communication. PIER Discussion Papers. 2017;54:1–42.
19. Debnath R, Bardhan R. India nudges to contain COVID-19 pandemic: A reactive public policy analysis using machine-learning based topic modelling. PLoS One. 2020;15(9):e0238972. https://doi.org/10.1371/journal.pone.0238972 PMID: 32915899
20. Blei D, Ng A, Jordan M. Latent dirichlet allocation. J Mach Learn Res. 2003;3(1):993–1022.
21. Xie Q, Zhang X, Ding Y, Song M. Monolingual and multilingual topic analysis using LDA and BERT embeddings. J Informetr. 2020;14(3):101055. https://doi.org/10.1016/j.joi.2020.101055
22. Box GEP, Pierce DA. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J Am Stat Assoc. 1970;65(332):1509–26. https://doi.org/10.1080/01621459.1970.10481180
23. Canagarajah S. Ethnographic methods in language policy. An introduction to language policy: theory and method. 2006.
24. Burstein P. The impact of public opinion on public policy: A review and an Agenda. Polit Res Q. 2003;56(1):29–40. https://doi.org/10.2307/3219881
25. Smits J, Gündüz-Hoşgör A. Linguistic capital: Language as a socio-economic resource among Kurdish and Arabic women in Turkey. Ethn Racial Stud. 2003;26(5):829–53. https://doi.org/10.1080/0141987032000109050
26. Blachford DR. Language planning and bilingual education for linguistic minorities in china, a case study of the policy formulation and implementation process. 1999.
27. Spolsky B. Language management in the People's Republic of China. Language. 2014;90(4):e165–79. https://doi.org/10.1353/lan.2014.0075
28. Baldauf RB Jr. Rearticulating the case for micro language planning in a language ecology context. Curr Issues Lang Plan. 2006;7(2–3):147–70. https://doi.org/10.2167/cilp092.0
29. Hopkins M. Beliefs in context t: Understanding language policy implementation at a systems level. Educ Policy. 2014;30(4):573–605. https://doi.org/10.1177/0895904814550073
30. Beukes AM. Language policy implementation in South Africa: how Kempton Park's great expectations are dashed in Tshwane. Stellenbosch Pap Linguist. 2008;38:1–26.
31. Shohamy E. Language policy and language assessment: The relationship. Curr Issues Lang Plan. 2008;9(3):363–73. https://doi.org/10.1080/14664200802139604
32. Dreyer JT. The evolution of language policies in China. 2003.
33. Hua Z, Wei L. Transnational experience, aspiration and family language policy. J Multiling Multicult Dev. 2016;37(7):655–66. https://doi.org/10.1080/01434632.2015.1127928

34. Huang G, Zhao R. Harmonious discourse analysis: approaching peoples' problems in a Chinese context. Lang Sci. 2021;85:101365. https://doi.org/10.1016/j.langsci.2021.101365
35. Majhanovich S. Neo-liberalism, globalization, language policy and practice issues in the Asia-Pacific region. Asia Pac J Educ. 2014;34(2):168–83. https://doi.org/10.1080/02188791.2013.875650
36. Grin F. Economic considerations in language policy. An introduction to language policy: theory and method. Publisher Name. 2006:77–94.
37. Shohamy E. Language policy: Hidden agendas and new approaches: Routledge. 2006.
38. Wickström B, Gazzola M. The economics of language policy and planning. The Routledge handbook of language policy and planning. London: Routledge. 2023:158–71.
39. Dunmore SS. Language policy and prospects: Metalinguistic discourses on social disruption and language maintenance in a transatlantic, minority community. Lang Commun. 2021;76:69–78. https://doi.org/10.1016/j.langcom.2020.10.006
40. Feldman R, Sanger J. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge Univ Press. 2006.
41. Liu Q, Jia M, Xia D. Dynamic evaluation of new energy vehicle policy based on text mining of PMC knowledge framework. J Clean Prod. 2023;392:136237. https://doi.org/10.1016/j.jclepro.2023.136237
42. Liu J. Lexical features of economic legal policy and news in china since the COVID-19 outbreak. Frontiers in Public Health. 2022;10:928965. https://doi.org/10.3389/fpubh.2022.928965 PMID: 35844862
43. Zong C, Xia R, Zhang J. Text data mining. Springer. 2021.
44. Agnihotri D, Verma K, Tripathi P. Pattern and cluster mining on text data. In: 2014 Fourth International Conference on Communication Systems and Network Technologies. IEEE. 2014:428–32. https://doi.org/10.1109/csnt.2014.92
45. ShrihariR C, Desai A. A review on knowledge discovery using text classification techniques in text mining. IJCA. 2015;111(6):12–15. https://doi.org/10.5120/19542-0784
46. Ramya R, Venugopal K, Iyengar S, Patnaik L. Feature extraction and duplicate detection for text mining: a survey. Glob J Comput Sci Technol. 2017;16(5):1–21.
47. Zhai C, Massung S. Text data management and analysis: a practical introduction to information retrieval and text mining. Association for Computing Machinery and Morgan & Claypool. 2016.
48. Yadollahi A, Shahraki AG, Zaiane OR. Current state of text sentiment analysis from opinion to emotion mining. ACM Comput Surv. 2017;50(2):1–33. https://doi.org/10.1145/3057270
49. Galati F, Bigliardi B. Industry 4.0: Emerging themes and future research avenues using a text mining approach. Comput Ind. 2019;109:100–13. https://doi.org/10.1016/j.compind.2019.04.018
50. Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. Int J Med Inform. 2019;125:37–46. https://doi.org/10.1016/j.ijmedinf.2019.02.008 PMID: 30914179
51. Yi J, Nasukawa T, Bunescu R, Niblack W. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In: Third IEEE International Conference on Data Mining. IEEE Comput. Soc. 2003:427–34. https://doi.org/10.1109/icdm.2003.1250949
52. Tollefson J. Language policies in education. London: Routledge. 2012.
53. Christensen G, Stanat P. Language policies and practices for helping immigrants and second-generation students succeed. The Transatlantic Taskforce on Immigration and Integration, Migration Policy Institute and Bertelsmann Stiftung. 2007.
54. May S. Deconstructing the instrumental/identity divide in language policy debates. 2005.
55. Lempert RJ. Shaping the next one hundred years: new methods for quantitative, long-term policy analysis. 2003.
56. Oxley J, Günhan E, Kaniamattam M, Damico J. Multilingual issues in qualitative research. Clin Linguist Phon. 2017;31(7–9):612–30. https://doi.org/10.1080/02699206.2017.1302512 PMID: 28665758
57. Zhou L, Pan S, Wang J, Vasilakos AV. Machine learning on big data: Opportunities and challenges. Neurocomputing. 2017;237:350–61. https://doi.org/10.1016/j.neucom.2017.01.026
58. Hassani H, Beneki C, Unger S, Mazinani MT, Yeganegi MR. Text mining in big data analytics. BDCC. 2020;4(1):1. https://doi.org/10.3390/bdcc4010001
59. Zhang Y, Zhang G, Chen H, Porter AL, Zhu D, Lu J. Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research. Technol Forecast Soc Change. 2016;105:179–91. https://doi.org/10.1016/j.techfore.2016.01.015
60. Eddamiri S, Benghabrit A, Zemmouri E. RDF graph mining for cluster-based theme identification. IJWIS. 2020;16(2):223–47. https://doi.org/10.1108/ijwis-10-2019-0048
61. Vaismoradi M, Jones J, Turunen H, Snelgrove S. Theme development in qualitative content analysis and thematic analysis. J Adv Nurs. 2016.
62. Chabi A, Kboubi F, Ahmed M. Thematic analysis and visualization of textual corpus. 2011.
63. Guo L, Vargo CJ, Pan Z, Ding W, Ishwar P. Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. J Mass Commun Q. 2016;93(2):332–59. https://doi.org/10.1177/1077699016639231

64. Chizari H, Ismail F. A grid-insensitive LDA method on triangular grids solving the system of euler equations. J Sci Comput. 2016;71(2):839–74. https://doi.org/10.1007/s10915-016-0323-5
65. Liu B. Sentiment analysis: mining opinions, sentiments, and emotions. Cambridge University Press. 2020.
66. Maçaira PM, Tavares Thomé AM, Cyrino Oliveira FL, Carvalho Ferrer AL. Time series analysis with explanatory variables: A systematic literature review. Environ Model Softw. 2018;107:199–209. https://doi.org/10.1016/j.envsoft.2018.06.004
67. Sapankevych N, Sankar R. Time Series Prediction Using Support Vector Machines: A Survey. IEEE Comput Intell Mag. 2009;4(2):24–38. https://doi.org/10.1109/mci.2009.932254
68. Çalişir D, Doğantekin E. An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier. Expert Syst Appl. 2011;38(7):8311–5. https://doi.org/10.1016/j.eswa.2011.01.017
69. Chauhan U, Shah A. Topic Modeling Using Latent Dirichlet allocation. ACM Comput Surv. 2021;54(7):1–35. https://doi.org/10.1145/3462478
70. Alsmearat K, Al-Ayyoub M, Al-Shalabi R. An extensive study of the Bag-of-Words approach for gender identification of Arabic articles. In: 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA). IEEE. 2014:601–8. https://doi.org/10.1109/aiccsa.2014.7073254
71. Tuarob S, Pouchard LC, Giles CL. Automatic tag recommendation for metadata annotation using probabilistic topic modeling. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries. 2013.
72. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl. 2018;78(11):15169–211. https://doi.org/10.1007/s11042-018-6894-4
73. Blei D, Ng A, Jordan M. Latent dirichlet allocation. In: Adv Neural Inf Process Syst. 2001.
74. Gan J, Qi Y. Selection of the optimal number of topics for LDA topic model-taking patent policy analysis as an example. Entropy (Basel). 2021;23(10):1301. https://doi.org/10.3390/e23101301 PMID: 34682025
75. Siami-Namini S, Namin AS. Forecasting economics and financial time series: ARIMA vs. LSTM. In: arXiv preprint. 2018.
76. Ariyo AA, Adewumi AO, Ayo CK. Stock price prediction using the ARIMA model. In: 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation. IEEE. 2014:106–12. https://doi.org/10.1109/uksim.2014.67
77. Bazrafshan O, Salajegheh A, Bazrafshan J, Mahdavi M, Fatehi Maraj A. Hydrological drought forecasting using ARIMA models (case study: Karkheh basin). Ecopersia. 2015;3(3):1099–117.
78. Ordóñez C, Sánchez Lasheras F, Roca-Pardiñas J, Juez FJ de C. A hybrid ARIMA–SVM model for the study of the remaining useful life of aircraft engines. J Comput Appl Math. 2019;346:184–91. https://doi.org/10.1016/j.cam.2018.07.008
79. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr. 1974;19(6):716–23. https://doi.org/10.1109/tac.1974.1100705
80. Schwarz G. Estimating the dimension of a model. Ann Statist. 1978;6(2):461–4. https://doi.org/10.1214/aos/1176344136
81. Box G, Jenkins G, Reinsel G, Ljung G. Time series analysis: forecasting and control. John Wiley & Sons. 2015.
82. Shumway R, Stoffer D. Time series analysis and its applications. Springer. 2000.
83. Hasan M, Rahman A, Karim M, Khan M, Islam M. Normalized approach to find optimal number of topics in latent dirichlet allocation (lda). In: Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE. Springer Singapore. 2020:341–54.
84. Newman D, Lau J, Grieser K, Baldwin T. Automatic evaluation of topic coherence. In: Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics. 2010.
85. Mushtaq R. Augmented dickey fuller test. 2011.
86. As'ad M. Finding the best ARIMA model to forecast daily peak electricity demand. 2012.
87. Yamak P, Yujian L, Gadosey P. A comparison between arima, lstm, and gru for time series forecasting. In: Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence. 2019.
88. Smagulova J. Language policies of kazakhization and their influence on language attitudes and use. Int J Biling Educ Biling. 2008;11(3–4):440–75. https://doi.org/10.1080/13670050802148798
89. Ricento T. An introduction to language policy: theory and method. John Wiley & Sons. 2005.
90. Johnson D, Johnson D. What is language policy?. Springer. 2013.
91. Massey AK, Eisenstein J, Anton AI, Swire PP. Automated text mining for requirements analysis of policy documents. In: 2013 21st IEEE International Requirements Engineering Conference (RE). IEEE. 2013:4–13. https://doi.org/10.1109/re.2013.6636700
92. チョウアルバート. Envisioning Chinese as a global language. Glob Stud. 2022;6:55–70.
93. Yamat H, Umar NFM, Mahmood MI. Upholding the Malay Language and strengthening the English language policy: An education reform. IES. 2014;7(13):197–205. https://doi.org/10.5539/ies.v7n13p197
94. Xu H, Zhou H, Xu Y. Development of educational attainment and gender equality in China: new evidence from the 7th National Census. China Popul Dev Stud. 2022;6(4):425–51. https://doi.org/10.1007/s42379-022-00122-z

95. Hu G. English language education in China: Policies, progress, and problems. Lang Policy. 2005;4(1):5–24. https://doi.org/10.1007/s10993-004-6561-7
96. Chernysh VV, Vaseiko Y, Kaplinskiy V, Tkachenko L, Bereziuk J. Modern methods of training foreign language teachers. IJHE. 2020;9(7):332–44. https://doi.org/10.5430/ijhe.v9n7p332
97. Yuming L. Theories and practices of China's language resources. Cociolinguistics. 2020;3(3):10–29.
98. Zhang H, Cai S. Putonghua vs. minority languages: distribution of language laws, regulations, and documents in mainland China. Ethn Racial Stud. 2020;44(14):2574–94. https://doi.org/10.1080/01419870.2020.1828598
99. Wu X. From assimilation to autonomy: Realizing ethnic minority rights in China's national autonomous regions. Chin J Int Law. 2014;13(1):55–90. https://doi.org/10.1093/chinesejil/jmu006
100. Fan J, Gao Y, Zhao N, Dai R, Zhang H, Feng X, et al. Bibliometric analysis on COVID-19: A comparison of research between English and Chinese studies. Front Public Health. 2020;8:477. https://doi.org/10.3389/fpubh.2020.00477 PMID: 32923422
101. Wang D, Ding N, Li P, Zheng H-T. Cline: Contrastive learning with semantic negative examples for natural language understanding. 2021.
102. Han Y, Wu X. Language policy, linguistic landscape and residents' perception in Guangzhou, China: dissents and conflicts. Curr Issues Lang Plan. 2019;21(3):229–53. https://doi.org/10.1080/14664208.2019.1582943
103. Hartig F. A decade of wielding soft power through Confucius institutes: some interim results. Soft power with Chinese characteristics. New York: Routledge. 2019:133–47.