


Religious Policy in the TAR, 2014–24: Topic Modelling a Tibetan-Language Corpus with BERTopic

Ronald Schwartz (Memorial University of Newfoundland)
and
Robert Barnett (SOAS University of London)

he study of discourses, in the sense of narratives or themes, is essentially an historical project: a discourse generally has a life-span of some sort, emerging at a certain time, reaching a degree of prominence or pervasiveness, and then, in most cases, fading away or shifting into some other form. In this paper, we look at a set of computational tools that can be developed for tracing the epihistorical footprints left by certain discourses and discuss their use for the analysis and detection of such histories in modern Tibetan texts. Ultimately, these tools will be useable with any Tibetan texts that have been digitised, such as the newspapers from the 1950s and 1960s that are the target of the Divergent Discourses project.¹ For the initial development and testing of these tools, however, we used a set of texts in modern Tibetan that are “born-digital” – that is, they are already available in a digital format and, therefore, do not need to be photographed, scanned or machine-read.

The texts that we used for this study are articles taken from the premier Tibetan-language newspaper in the People’s Republic of

¹ The project received funding from the Deutsche Forschungsgemeinschaft (DFG) under project number 508232945 (<https://gepris.dfg.de/gepris/projekt/508232945?language=en>), and from the Arts and Humanities Research Council (AHRC) under project reference AH/X001504/1 (<https://gtr.ukri.org/projects?ref=AH%2FX001504%2F1>). For more information on Divergent Discourses, see <https://research.uni-leipzig.de/diverge/>.

China (PRC), *Tibet Daily*, known in Chinese as *Xizangribao* (西藏日报) and in Tibetan as the *Bod ljongs nyin re'i tshags par* (བོད་ལྗོངས་ཉིན་རེའི་ཚགས་པར་). *Tibet Daily* is not a newspaper in the normal sense of the word, where the principal purpose is to disseminate news. Rather, it is what might elsewhere be known as a gazetteer, in that it serves primarily to publicise decisions, projects, goals and opinions held by its proprietary institution: the Committee of the Tibet Autonomous Region (TAR) Branch of the Chinese Communist Party (CCP). As a result, we can, as analysts, assume that in most cases, a “discourse” or topic found in the pages of *Tibet Daily* will reflect not one or other commonly held opinion or narrative circulating organically within society, but a project, policy, opinion or goal that the TAR branch of the CCP intends at that time to publicise or implement.

In most cases, the presence of such a discourse in the columns of *Tibet Daily* will indicate that the CCP has embarked upon an organised effort to disseminate a specific belief or practice among the public. These efforts will reflect the initiation of a policy or political program, which at times will take the particularly intense form of political or social mobilisation known in Chinese as a campaign (运动 *yundong*; Tib. *las 'gul*), when Party officials, government workers and Party activists will be sent throughout the region to achieve the goals of that campaign. Consequently, a computational tool that can trace the historical arc of a discourse or topic can, when applied to a publication such as the *Tibet Daily*, be used to identify the beginnings, peaks and endings of the political campaigns, drives, policies and official opinions that shape, or attempt to shape, much of public and private life in contemporary Tibet. This paper describes our development of a dynamic topic modelling tool for modern Tibetan texts that can help analysts trace the rise and fall of such campaigns, policies and opinions in Tibet over time.

1 Creating the *Tibet Daily* corpus

Tibet Daily began publication in 1956. Nearly a year earlier, in October 1955, China's then leader, Mao Zedong (1893–1976), had sent a

message to Zhang Jingwu (张经武, 1906–1971), the Chinese government’s leading representative in Tibet, endorsing the decision to produce a newspaper in Tibet that would serve as the official organ of the CCP in Tibet. Reportedly, Mao Zedong wrote to Zhang:

"When running a newspaper in a minority area, the first thing to do is to run a newspaper in the minority language ... Unlike Qinghai, Tibet should not have a newspaper which is [bilingual] in Tibetan and Chinese, but [it should have one that is only] in Tibetan. The name of the newspaper and how it should be run should be discussed with the Tibetan authorities, who should decide, and we should not take over the running of the newspaper."²

This advice, emblematic of the concessional, co-optive approach of the CCP leadership towards central Tibetans at the time, was taken up by Zhang and his colleagues, at least on the surface: the new paper was published in two separate editions, one in Chinese and another in Tibetan, and Tibetan leaders were included in the decision as to its name. The paper rapidly became a major institution in the region, representing a core function of the new Chinese administration in Tibet, with a staff of 350 people by the early 2000s (Zhang 2004: 141). The political importance of the paper was underlined by the fact that it had to be defended by its own 60-strong militia unit during the uprising of March 1959, when it came under sustained, but ultimately unsuccessful, attack from Tibetan rebels (Zhao 1987: 179). Much the same happened in July 1966, when leftist activists besieged the *Tibet Daily* compound, this time in the name of Chairman Mao rather than

² “在少数民族地区办报，首先应办少数民族文字的报。” “西藏与青海不同，不要藏汉两文合版，要办藏文报。报纸用什么名字和怎样办好，应同西藏地方商量，由他们决定，我们不要包办。” (Dangdai Zhongguo congshu bianjibu [Editorial Board of the Series on Contemporary China], 当代中国的西藏 [Tibet in Contemporary China], vol. 2, Beijing: Contemporary China Press, 1991, p. 435). See <https://www.google.co.uk/books/edition/当代中国的西藏/UYm6AAAAIAAI> (accessed January 15, 2025); see also “*Tibet Daily*”, unsigned entry primarily authored by user “TinaLees-Jones”, Wikipedia, https://en.wikipedia.org/wiki/Tibet_Daily (accessed January 15, 2025).

Tibetan independence (Goldstein *et al.* 2009: 27–30; Tsering Woeser 2020: 372).

Although Mao is said to have emphasised the importance of the Tibetan-language edition of *Tibet Daily*, the Chinese-language version has come to have, and probably always had, a dominant role. By the mid-1990s, the print run for the Chinese edition was nearly 30,000, while the Tibetan edition reached 20,000; the number of contributors to the Chinese edition by that time was over 2,300, about twice the number who had written for the paper in Tibetan (Hartley 2005: 248). In addition, our research has found that, at least since 2014, there are substantially more articles in the Chinese-language edition of *Tibet Daily* than in the Tibetan-language edition. On average there are roughly twice as many articles in a given year in the Chinese edition compared to the Tibetan. Thus, in 2020, for instance, there were 6,295 articles in the Tibetan edition and 12,146 in the Chinese edition. More significantly, most Tibetan articles appeared to be translations of Chinese articles. We have not explored the differences between the two editions, but it is clear that the content we are primarily interested in — announcements of new policies and their implementation, important political meetings, speeches by leaders, slogans and political campaigns — seem to appear in the Chinese edition and then, usually a day or more later, are mirrored in the Tibetan version.

However, the Tibetan-language edition of *Tibet Daily* has played a significant role in promoting the vernacularisation of the language, according to Hartley (2005). Following the death of Mao and the end of the Cultural Revolution, it also made some contributions to the development of modern Tibetan literature, particularly by adding a literary column called *Smyug bsar* [New Pen] and, from the early 1980s onwards, publishing occasional short stories by Tibetan writers (Hartley 2005: 248).

For 40 years, the two editions of *Tibet Daily* were produced on paper, mainly for distribution in government offices and similar institutions. From the early 21st century, however, *Tibet Daily* began to appear online as well as in print. Currently, it is available in several online versions, with its main Chinese-language hub at <https://www.xzxw.com/> (accessed January 15, 2025; the site name is an acronym of

Xizangxinwen, “Tibet news”). This web portal aggregates multiple articles on different subjects published on different dates from various newspapers produced by the authorities in the TAR. These articles are accessed through links on pages that facilitate recursive navigation. This site has subsidiary pages which carry Tibetan-language and English-language versions of articles from *Tibet Daily* and other official publications, again in an aggregated format.

However, since 2008 *Tibet Daily* has also been available online in an e-paper format. This format reproduces the look of the printed edition, with a photographic image (actually, a PDF) of each page on the left of the screen and a text version of each article on the right. This static website has a flat structure, with content organised as a single daily edition with no recursive access to previous editions. The individual articles are HTML files organised chronologically by date, and thus it is possible to assemble a corpus covering the entire timespan of the e-paper website to the present. The Tibetan-language edition is accessed separately at a site with the root name <https://e.xzxw.com/xzrbzw/> (for Xizangribao Zangwen, “Tibetan-language Tibet Daily”), followed by the date of that day’s issue.

Articles from the Chinese-language edition of *Tibet Daily* are available in HTML format at <https://e.xzxw.com/xzrb/> (followed by the date) from 2008 onwards. However, the Tibetan-language site only has articles in HTML format from 2014. Before that, each issue of the e-paper is displayed as single whole pages in PDF format and thus is not easily recoverable as text. For the purpose of this study, we will therefore use the Tibetan-language edition of *Tibet Daily* from 2014 to 2024.

The articles we collected are all in HTML format and use a Unicode-compliant font for Tibetan characters. To facilitate analysis using the research tools developed for this study, the content of the articles was extracted using tags embedded in the HTML files. The paragraphs within the articles have been individually tagged, enabling the creation of a CSV file for each year of the corpus, with a row for every paragraph, along with metadata (article titles, dates, filename of original article). These CSV files comprise the corpus for research.

2 *Semantic Searching with Vector Embeddings*

This paper will demonstrate the use of tools that employ vector embeddings derived from transformer-based large language models (LLMs) to analyse Tibetan texts. Embeddings are numerical encodings that locate lengths of texts (phrase, sentences, paragraphs) in a high-dimensional vector space where semantically similar texts are close together in the vector space. Embeddings capture contextual meaning rather than mere word co-occurrences, providing a richer representation of language and meanings. There are a number of embedding models available for high-resource languages such as English or Chinese, but Tibetan is a relatively low-resource language where training has necessarily been limited to relatively small datasets. Ideally, multilingual embedding models will locate texts with the same or similar meanings in two or more different languages close to each other within the vector space.³

After examining several multilingual models, we found one that performs well with modern newspaper Tibetan — the version 2.0

³ There has been relatively little work to date using vector embeddings with Tibetan-language texts. Meelen (2022) reports using FastText (<https://fasttext.cc/>, accessed January 15, 2025) to generate word embeddings for classical Tibetan. Sabbagh (2023) uses the LASER multilingual sentence encoder from Facebook (<https://github.com/facebookresearch/>, accessed January 15, 2025) to align Tibetan translations of English language sentences. Neither of these models are transformer-based. Two transformer-based embedding models have been developed by researchers in the PRC using the BERT model — Tibetan-BERT from a team at Tibet University (https://huggingface.co/UTibetNLP/tibetan_bert, accessed January 15, 2025) and TiBERT from a team at Minzu University (<https://huggingface.co/CMLI-NLP/TiBERT>, accessed January 15, 2025). Both of these models are designed and trained for downstream tasks of text classification. BGE-m3 from the Beijing Academy of Artificial Intelligence is an open-source multilingual embedding model, designed for information retrieval applications, that offers Tibetan embeddings (<https://huggingface.co/BAAI/bge-m3>, accessed January 15, 2025). Amazon Web Services Titan text embeddings v2 model also includes Tibetan embeddings (<https://docs.aws.amazon.com/bedrock/latest/userguide/titan-embedding-models.html>, accessed January 15, 2025). Neither of these performed adequately for purposes of semantic searching and topic modelling. The newer Cohere version 3.0 model, which specialises in information retrieval tasks, also did not perform as well as version 2.0 (see Engels *et al.* 2025).

multilingual model from Cohere (<https://cohere.com/>, accessed January 15, 2025; see Engels & Barnett 2025 in this volume). The embeddings are optimised for multilingual text understanding and accessible through an API. We found that they also perform well with cross-lingual queries in Tibetan, Chinese, and English. This model has a 256-token context limit, which requires that the paragraphs must be “chunked” into phrases that fall within this limit before generating the embeddings. A routine is implemented to “chunk” or split the paragraphs at the last (!) *shad* (the Tibetan punctuation mark most often used to mark the end of a sentence or phrase) within 256 tokens. Vector embeddings are generated and aligned for every chunked paragraph using the Cohere API.⁴

Semantic searching is the basic tool for investigating the corpus. Topic modelling, which we describe later, relies on the same vector embeddings as semantic searching and applies the same measure of similarity. Using the Cohere embeddings we can search throughout the entire corpus of *Tibet Daily* from 2014 to 2024 for chunks of text (and corresponding paragraphs) that are semantically similar to our query text. The query can be just a few words or a phrase, but for exploring the corpus and identifying topics and themes, it is more effective to include a chunk of text drawn from the corpus that is representative of the content being searched for rather than an individual word or short string. The semantic search program (written in Python) calls the Cohere API to generate an embedding for the query and then compares the numerical encoding of the query with the previously generated encodings of every chunk in the entire corpus. The results are displayed in descending order of similarity. If

⁴ The Cohere embedding model uses a WordPiece tokeniser, which segments the texts into single Tibetan syllables and uses ## for subword units. The Tibetan *tsheg* is also treated as a token. An example of a tokenised text: [‘[CLS]’, ‘མ’, ‘##ར’, ‘’, ‘ཅ’, ‘##ུན’, ‘’, ‘།’, ‘##ེ’, ‘##ེང’, ‘’, ‘གིས’, ‘’, ‘ནན’, ‘’, ‘བཤད’, ‘’, ‘གནང’, ‘’, ‘དོན’, ‘།’, ‘ཟེ’, ‘’, ‘པ’, ‘’, ‘ལག’, ‘’, ‘དང’, ‘’, ‘ཟེ’, ‘’, ‘ཚོན’, ‘’, ‘ལག’, ‘’, ‘གིས’, ‘’, ‘དག’, ‘’, ‘ཟེ’, ‘##ར’, ‘’, ‘ལ’, ‘’, ‘དམ’, ‘’, ‘འཛོལ’, ‘’, ‘ཡག’, ‘’, ‘ཟོ’, ‘’, ‘ཅུ’, ‘’, ‘ཅུ’, ‘’, ‘ནི’, ‘’, ‘རང’, ‘’, ‘ལ’, ‘##ལ’, ‘##ེའི’, ‘##འི’, ‘’, ‘ལས’, ‘’, ‘འགན’, ‘’, ‘ཡིན’, ‘’, ‘པ’, ‘’, ‘དང’, ‘’, ‘།’, ‘དག’, ‘’, ‘ཟེ’, ‘##ར’, ‘’, ‘ལ’, ‘’, ‘དམ’, ‘’, ‘འཛོལ’, ‘’, ‘མ’, ‘’, ‘བྱས’, ‘’, ‘པ’, ‘’, ‘ནི’, ‘’, ‘འགན’, ‘’, ‘ཤོང’, ‘’, ‘ཡིན’, ‘’, ‘པ’, ‘’, ‘དག’, ‘’, ‘ཟེ’, ‘##ར’, ‘’, ‘ཡག’, ‘’, ‘ཟོ’, ‘’, ‘མ’, ‘’, ‘བྱས’, ‘’, ‘པ’, ‘’, ‘ནི’, ‘’, ‘འགན’, ‘’, ‘ལ’, ‘##ལ’, ‘##ེལ’, ‘’, ‘ཡིན’, ‘’, ‘པ’, ‘’, ‘བཅས’, ‘’, ‘།’, ‘’, ‘ལྟ’, ‘’, ‘གིས’, ‘’, ‘བཤད’, ‘’, ‘ཟོ’, ‘’, ‘བརྟུགས’, ‘’, ‘ནས’, ‘།’, ‘[SEP]’].

the query is itself a chunk from the corpus, it will be displayed first in the list of results and will have the highest similarity value. The program allows the user to specify the number of hits to return. For our research purposes, we might ask it initially to return as many as fifty or one hundred hits in descending order of similarity to get a sense of the scope of a query, as we are interested not just in the few top-most similar hits, but in exploring similar examples of discourse in many paragraphs in many articles published at different times.

Table 1 is an example of the results from a query using semantic search. Just the first four rows from a search are displayed here to illustrate the use of the search tool. In the first row, the query (a selection of text from the corpus) returns itself. The next three rows are paragraphs in descending order of similarity to the query. In this example, the query is used to locate material in the corpus that mentions implementing the “four standards” drive in monasteries and nunneries (one of the topics that will be discovered through topic modelling). Having found an instance of the “four standards” in one context in the corpus, semantic search makes it possible to find other contexts as well: #1 refers to a high-level meeting of officials and religious leaders in which the drive is discussed; #2 refers to the implementation of “four standards” education in Shigatse; and in #3, Ding Yexian, a Deputy Party Secretary of the TAR branch of the CCP, compliments the monasteries of Drepung and Sera for their implementation of “four standards” education.

Semantic searching can be used along with topic modelling to identify and explore discourses within a corpus. Having located articles with relevant paragraphs makes it easy to go directly to the article. The search program also returns the entire paragraph for each chunk. Metadata is prepended to the displayed text that identifies the original *Tibet Daily* article (an HTML file) in the corpus where the chunk is located and its publication date. A unique number is assigned to every paragraph in the entire corpus. The search program also returns and displays the full paragraph from which the chunk is taken. A measure of similarity to the query text is also shown (the Cohere version 2.0 multilingual embedding model uses the inner dot product

to measure similarity; unlike cosine similarity, this measure of similarity is not normalised and can be larger than 1).

Translations from Tibetan to English of the paragraphs are provided by the Azure multilingual translation API.⁵ There are now several translators available for modern Tibetan. We have found that the Azure translator API (version 3.0) currently provides the best translation of modern newspaper Tibetan, though it still makes both syntactical and lexical mistakes.

Table 1: Semantic Search

No.	Chunk	Original Full Paragraph	Similarity
Qry.	323283.01 content_853060.htm 2018-09-12 རང་ལྗོངས་ཀྱིས་“ཚད་གཞི་ལག་བཞེར་བཅུ་གཉིས་། རྩོད་ཐོན་གྱི་ བཙུན་ཏུར་ཐག་ཉེད་”ཅེས་པའི་སློབ་གསོ་ལག་ལེན་ཉེད་སློབ་སློབ་ཚུན་། ས་གནས་ལག་གིས་ཚད་མཐོའི་མཐོང་ཚེན་བྱས་པ་དང་། རྩོད་ཚན་ལག་ གིས་ལས་ཀ་ཏུར་ཐག་བསྐྱབས་ཤིང་། དགོན་སྡེའི་གྲྭ་བཙུན་གྱིས་སློབ་ འགྲུལ་ལ་གཞོགས་འདེགས་ཏུར་ཐག་དང་། སློབ་སློབ་ལ་རང་འགྲུལ་ རང་ལྗོངས་པ། རང་རྩོགས་རང་སྲོང་ཚོར་འབྲི་བ་བཅས་བྱས་ཏེ་ལྗོངས་ ཡོངས་ཀྱི་གྲྭ་བཙུན་གྱི་བསམ་སློབ་འཛིན་གཅིག་བྱུར་དང་། བོད་ བརྒྱུད་ནང་བསྟན་གྱི་དགོན་སྡེའི་རྒྱན་གཏེན་སློབ་སློབ་ལ་སླུང་སློང་། ཚོས་ ལུགས་ལྟུང་ལོངས་རྒྱན་མཐུན་བཅུད་བཅུད་ལྟེང་དང་། ཡུན་རིང་བཅུན་སྡེང་། ལྟོན་ཡོངས་བཅུན་སྡེང་བཅས་འགན་ལེན་བྱུང་ཁར་སློབ་གསོ་ལག་ལེན་ ཉེད་སློབ་སློབ་འབྲས་བུ་ལྟར་དུ་བ་ཐོབ་ཡོད། འདི་ག་ཚགས་པར་ཐོག་ དེ་རིང་ནས་བཟུང་“ཚད་གཞི་ལག་བཞེར་བཅུ་གཉིས་། རྩོད་ཐོན་གྱི་ བཙུན་ཏུར་ཐག་ཉེད་”ཅེས་པའི་ཚད་སློབ་ལེ་ཚན་འདོན་རྒྱ་ཡིན་པས། དོ་ ལུར་ཡོད་པ་ལྟ།	191.53	

⁵ <https://azure.microsoft.com/en-us/services/cognitive-services/translator/> (accessed January 15, 2025).

No.	Chunk	Original Full Paragraph	Similarity
1	<p>616415.01 content_121540.html 2021-12-31</p> <p>རང་སྐོར་ལྷོངས་ཏང་ལུད་ཀྱི་རྒྱན་ལུ་འཐབ་ འཐབ་ལུ་གཅིག་ལྷོངས་ལུ་ཡི་ཡུ་ཀྱང་ཀམ་ཚོ་བརྟན་ ལྱིས་“ཚད་གཞི་བཞེར་བཅུ་སྤྲོད་གིས་སྤྲོད་ཐོན་གྱ་ བཅུན་ཏུར་ཐག་བྱེད་པའི་”ལྷོངས་ཡོངས་ཀྱི་སློབ་གསོ་ལག་ལེན་བྱེད་སློ་ སྤྲོད་ལྷོངས་ལུ་གནས་ཚུལ་སྤྲོད་མེད་ཞུས་པ་རེད། ཚོས་ལུགས་ལས་ རིགས་ཀྱི་འཕུལ་མི་སྤྲོད་ལུ་འཐབ་བསྟན་མཁས་ལྷོངས་ལྷོངས་དང་། བཀྲ་ཤིས་ རྒྱལ་མཚན། ཀྱམ་བཟང་དབང་འདུས། ལླ་བ་ཚེ་རིང་། རྩ་ཐོག་སློབ་བཟང་ ཡེ་ཤེས། རྗེ་རུང་བསྟན་པའི་རྒྱལ་མཚན། རྣམ་རྒྱལ་དབང་ལྷོངས་སློབ་བཟང་ བསམ་གཏམ་བཅས་ཀྱིས་སྤྲོད་པའི་ལོ་རྒྱུ་ལྷོངས་ལྷོངས་ལྷོངས་ལྷོངས་ལྷོངས་ བསྟན་མཁས་ལྷོངས་དང་། བཀྲ་ཤིས་རྒྱལ་ མཚན། ཀྱམ་བཟང་དབང་འདུས། ལླ་བ་ཚེ་ རིང་། རྩ་ཐོག་སློབ་བཟང་ཡེ་ཤེས། རྗེ་རུང་བསྟན་ པའི་རྒྱལ་མཚན། རྣམ་རྒྱལ་དབང་ལྷོངས་ སློབ་བཟང་བསམ་གཏམ་བཅས་ཀྱིས་སྤྲོད་པའི་ གཏམ་བཤད་གནང་ཞེས། ཚོས་མས་རང་ཉིད་ ཀྱི་དོན་དོམས་དང་རྒྱུང་འབྲེལ་བྱས་ཏེ།</p>	<p>616415 content_121540.html 2021-12-31</p> <p>རང་སྐོར་ལྷོངས་ཏང་ལུད་ཀྱི་རྒྱན་ལུ་འཐབ་འཐབ་ལུ་གཅིག་ལྷོངས་ལུ་ཡི་ཡུ་ ཀྱང་ཀམ་ཚོ་བརྟན་ལྱིས་“ཚད་གཞི་བཞེར་བཅུ་སྤྲོད་གིས་སྤྲོད་ཐོན་གྱ་ བཅུན་ཏུར་ཐག་བྱེད་པའི་”ལྷོངས་ཡོངས་ཀྱི་སློབ་གསོ་ལག་ལེན་བྱེད་སློ་ སྤྲོད་ལྷོངས་ལུ་གནས་ཚུལ་སྤྲོད་མེད་ཞུས་པ་རེད། ཚོས་ལུགས་ལས་ རིགས་ཀྱི་འཕུལ་མི་སྤྲོད་ལུ་འཐབ་བསྟན་མཁས་ལྷོངས་ལྷོངས་དང་། བཀྲ་ཤིས་ རྒྱལ་མཚན། ཀྱམ་བཟང་དབང་འདུས། ལླ་བ་ཚེ་རིང་། རྩ་ཐོག་སློབ་བཟང་ ཡེ་ཤེས། རྗེ་རུང་བསྟན་པའི་རྒྱལ་མཚན། རྣམ་རྒྱལ་དབང་ལྷོངས་སློབ་བཟང་ བསམ་གཏམ་བཅས་ཀྱིས་སྤྲོད་པའི་ལོ་རྒྱུ་ལྷོངས་ལྷོངས་ལྷོངས་ལྷོངས་ལྷོངས་ བསྟན་མཁས་ལྷོངས་དང་། བཀྲ་ཤིས་རྒྱལ་ མཚན། ཀྱམ་བཟང་དབང་འདུས། ལླ་བ་ཚེ་ རིང་། རྩ་ཐོག་སློབ་བཟང་ཡེ་ཤེས། རྗེ་རུང་བསྟན་ པའི་རྒྱལ་མཚན། རྣམ་རྒྱལ་དབང་ལྷོངས་ སློབ་བཟང་བསམ་གཏམ་བཅས་ཀྱིས་སྤྲོད་པའི་ གཏམ་བཤད་གནང་ཞེས། ཚོས་མས་རང་ཉིད་ ཀྱི་དོན་དོམས་དང་རྒྱུང་འབྲེལ་བྱས་ཏེ།</p>	187.09
2	<p>329616.01 content_858688.htm 2018-10-20</p> <p>བསྟན་ལོས་ནན་བཤད་གནང་དོན། རང་སྐོར་ ལྷོངས་ཏང་ལུད་དང་མིད་གཞུང་གིས་ལྷོངས་ ཡོངས་ཀྱི་ཚོས་ལུགས་ལུ་འཐབ་ལུ་ཚད་ གཞི་བཞེར་བཅུ་སྤྲོད་གིས་སྤྲོད་ཐོན་གྱ་བཅུན་ཏུར་ ཐག་བྱེད་པའི་”སློབ་གསོ་ལག་ལེན་བྱེད་སློ་ སྤྲོད་ལྷོངས་ལུ་གནས་ཚུལ་སྤྲོད་མེད་ཞུས་པ་རེད། ཚོས་ལུགས་ལས་ རིགས་ཀྱི་འཕུལ་མི་སྤྲོད་ལུ་འཐབ་བསྟན་མཁས་ ལྷོངས་ལྷོངས་དང་། བཀྲ་ཤིས་རྒྱལ་ མཚན། ཀྱམ་བཟང་དབང་འདུས། ལླ་བ་ཚེ་ རིང་། རྩ་ཐོག་སློབ་བཟང་ཡེ་ཤེས། རྗེ་ རུང་བསྟན་པའི་རྒྱལ་མཚན། རྣམ་རྒྱལ་ དབང་ལྷོངས་སློབ་བཟང་བསམ་གཏམ་བཅས་ ཀྱིས་སྤྲོད་པའི་གཏམ་བཤད་གནང་ཞེས། ཚོས་ མས་རང་ཉིད་ཀྱི་དོན་དོམས་དང་རྒྱུང་ འབྲེལ་བྱས་ཏེ།</p>	<p>329616 content_858688.htm 2018-10-20</p> <p>བསྟན་ལོས་ནན་བཤད་གནང་དོན། རང་སྐོར་ལྷོངས་ཏང་ལུད་དང་མིད་ གཞུང་གིས་ལྷོངས་ཡོངས་ཀྱི་ཚོས་ལུགས་ལུ་འཐབ་ལུ་ཚད་གཞི་བཞེར་ བཅུ་སྤྲོད་གིས་སྤྲོད་ཐོན་གྱ་བཅུན་ཏུར་ཐག་བྱེད་པའི་”སློབ་གསོ་ ལག་ལེན་བྱེད་སློ་སྤྲོད་ལྷོངས་ལུ་གནས་ཚུལ་སྤྲོད་མེད་ཞུས་པ་རེད། ཚོས་ ལུགས་ལས་རིགས་ཀྱི་འཕུལ་མི་སྤྲོད་ལུ་འཐབ་བསྟན་མཁས་ལྷོངས་ལྷོངས་ དང་། བཀྲ་ཤིས་རྒྱལ་མཚན། ཀྱམ་བཟང་དབང་འདུས། ལླ་བ་ཚེ་རིང་། རྩ་ ཐོག་སློབ་བཟང་ཡེ་ཤེས། རྗེ་རུང་བསྟན་པའི་རྒྱལ་མཚན། རྣམ་རྒྱལ་ དབང་ལྷོངས་སློབ་བཟང་བསམ་གཏམ་བཅས་ཀྱིས་སྤྲོད་པའི་གཏམ་བཤད་ གནང་ཞེས། ཚོས་མས་རང་ཉིད་ཀྱི་དོན་དོམས་དང་རྒྱུང་འབྲེལ་བྱས་ ཏེ།</p>	185.76

No.	Chunk	Original Full Paragraph	Similarity
3	329962.01 content_859093.htm 2018-10-23 རྒྱུ་ལེ་ཤུན་གྱིས་འབྲས་སྤྲེལ་དགོན་དང་མེ་ར་དགོན་གྱི་ཚད་གཞི་ དགོན་གྱི་ཚད་གཞི་བཞི་བརྗེས་ལག་བསྟར་ གྱིས་སྤྱོད་ཐོན་གྱ་བརྩམས་ཏུ་ཐག་བྱེད་པའི་” སྤྱོད་གསོ་ལག་ལེན་བྱེད་སྤྱོད་སྤྱི་ལ་བའི་ཐད་ཐོབ་ པའི་གྲུབ་འབྲས་ལ་གཤེད་འཛོག་གང་ལེགས་ གནང་ཞིང་། ཁོང་གིས་དགོན་སྡེ་གཉིས་གྱིས་ ལྷེ་བའི་ལས་དོན་ལ་དམ་པོར་དམིགས་པ་དང་། དམིགས་ཚད་དང་ལས་འགན་ལ་དམ་འཛིན་ ནན་པོ་བྱེད་པ། ལས་ཀྱི་བྱ་བ་ལས་འགས་ལ་གཏོད་ བྱེད་པ། བྱེད་སྤྱོད་གསོ་ལ་ལྷན་པའི་དང་སྤྱི་ལ་བ། མི་དང་། དེའི་ ལྷེ་བའི་ལས་དོན་ལ་དམ་པོར་དམིགས་པ་དང་། དམིགས་ཚད་དང་ལས་འགན་ལ་དམ་འཛིན་ ནན་པོ་བྱེད་པ། ལས་ཀྱི་བྱ་བ་ལས་འགས་ལ་གཏོད་ བྱེད་པ། བྱེད་སྤྱོད་གསོ་ལ་ལྷན་པའི་དང་སྤྱི་ལ་བ། མི་དང་། དེའི་ ལྷེ་བའི་ལས་དོན་ལ་དམ་པོར་དམིགས་པ་དང་། དམིགས་ཚད་དང་ལས་འགན་ལ་དམ་འཛིན་ ནན་པོ་བྱེད་པ། ལས་ཀྱི་བྱ་བ་ལས་འགས་ལ་གཏོད་ བྱེད་པ། བྱེད་སྤྱོད་གསོ་ལ་ལྷན་པའི་དང་སྤྱི་ལ་བ། མི་དང་། དེའི་	329962 content_859093.htm 2018-10-23 རྒྱུ་ལེ་ཤུན་གྱིས་འབྲས་སྤྲེལ་དགོན་དང་མེ་ར་དགོན་གྱི་ཚད་གཞི་ བཞི་བརྗེས་ལག་བསྟར་གྱིས་སྤྱོད་ཐོན་གྱ་བརྩམས་ཏུ་ཐག་བྱེད་ པའི་”སྤྱོད་གསོ་ལག་ལེན་བྱེད་སྤྱོད་སྤྱི་ལ་བའི་ཐད་ཐོབ་པའི་གྲུབ་འབྲས་ལ་ གཤེད་འཛོག་གང་ལེགས་གནང་ཞིང་། ཁོང་གིས་དགོན་སྡེ་གཉིས་གྱིས་ ལྷེ་བའི་ལས་དོན་ལ་དམ་པོར་དམིགས་པ་དང་། དམིགས་ཚད་དང་ལས་ འགན་ལ་དམ་འཛིན་ནན་པོ་བྱེད་པ། ལས་ཀྱི་བྱ་བ་ལས་འགས་ལ་གཏོད་ བྱེད་པ། བྱེད་སྤྱོད་གསོ་ལ་ལྷན་པའི་དང་སྤྱི་ལ་བ། མི་དང་། དེའི་ ལྷེ་བའི་ལས་དོན་ལ་དམ་པོར་དམིགས་པ་དང་། དམིགས་ཚད་དང་ལས་ འགན་ལ་དམ་འཛིན་ནན་པོ་བྱེད་པ། ལས་ཀྱི་བྱ་བ་ལས་འགས་ལ་གཏོད་ བྱེད་པ། བྱེད་སྤྱོད་གསོ་ལ་ལྷན་པའི་དང་སྤྱི་ལ་བ། མི་དང་། དེའི་ ལྷེ་བའི་ལས་དོན་ལ་དམ་པོར་དམིགས་པ་དང་། དམིགས་ཚད་དང་ལས་ འགན་ལ་དམ་འཛིན་ནན་པོ་བྱེད་པ། ལས་ཀྱི་བྱ་བ་ལས་འགས་ལ་གཏོད་ བྱེད་པ། བྱེད་སྤྱོད་གསོ་ལ་ལྷན་པའི་དང་སྤྱི་ལ་བ། མི་དང་། དེའི་ ལྷེ་བའི་ལས་དོན་ལ་དམ་པོར་དམིགས་པ་དང་། དམིགས་ཚད་དང་ལས་ འགན་ལ་དམ་འཛིན་ནན་པོ་བྱེད་པ། ལས་ཀྱི་བྱ་བ་ལས་འགས་ལ་གཏོད་ བྱེད་པ། བྱེད་སྤྱོད་གསོ་ལ་ལྷན་པའི་དང་སྤྱི་ལ་བ། མི་དང་། དེའི་	185.44

Table 1: Semantic Search (cont)

No.	Machine-translated Full Paragraph
Qry.	323283 content_853060.htm 2018-09-12 Since our district launched the educational practice of "abiding by the four standards and actively being advanced monks and nuns", all localities have attached great importance to it, and all units should take action. Actively acting, the monks and nuns of the temple actively cooperate with the preaching, take the initiative to participate in learning, consciously write their experiences, and unify the ideological understanding of the monks and nuns in the whole region. Maintain the normal order of Tibetan Buddhist monasteries, ensure sustained and long-term stability in the religious field, and achieve overall stability, and achieve results in educational and practical activities. Starting today, this newspaper will broadcast a column entitled "Abide by the Four Standards and Actively Become Advanced Monks and Nuns." Welcome to pay attention to it.
1	616415 content_121540.html 2021-12-31 Karma Tsedan, Member of the Standing Committee of the Party Committee and Minister of the United Front Work Department of the Autonomous Region, Presents the Educational Practice Activities of the Autonomous Region on "Complying with the Four Standards and Actively Being Advanced Monks and Nuns" Representatives of religious circles Zhukang Thubten Kedrup, Tashi Gangcun, Gongde Wangdui, Dawa Tsering, Mama Lobsang Yeshe, Jalen Tenzin Jebu and Langjie Wangdui. Luosang Jiangcun made an exchange speech, and in the light of their own realities, they talked freely about the results and experiences of participating in the educational practice activities, and guided the religious figures to love the party and the motherland." resolutely support the people's leader, the socialist system and

No.	Machine-translated Full Paragraph
	the system of regional ethnic autonomy, take a clear-cut stand against separatism, and safeguard the unity of the motherland and ethnic unity; It fully embodies the self-confidence and determination to listen to the party, feel the party's kindness, follow the party, continuously enhance the "five identities", firmly establish the sense of community of the Chinese nation, and promote the sinicisation of Tibetan Buddhism.
2	329616 content_858688.htm 2018-10-20 The party committee and government of the autonomous region decided to carry out the educational and practical activities of "abiding by the four standards and actively becoming advanced monks and nuns" in the religious field of the whole region. It is necessary to comprehensively and thoroughly study and implement Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era, and understand the spirit of the 19th National Congress of the Communist Party of China. All departments at all levels in Shigatse City should be of great strategic significance to carrying out educational practice activities, and earnestly strengthen the implementation of the strategic deployment of "one belt and one first". It is necessary to deeply understand, link up and down, cohesion, and everyone's participation, so as to stimulate the enthusiasm and initiative of monks and nuns to participate in educational practice activities. We will continue to develop the educational practice of "abiding by the four standards and becoming advanced monks and nuns" in depth.
3	329962 content_859093.htm 2018-10-23 Ding Yexian fully affirmed the achievements of Drepung Monastery and Sera Monastery in the educational practice of "abiding by the four standards and actively being advanced monks and nuns". He stressed that the two sessions closely focused on the central work, vigorously grasped the goals and tasks, innovated work methods, enlivened activities, and adhered to people, things, things, and things. It has played a leading role in building a team of monks and nuns who are "politically reliable, religiously high-level, morally fair, and effective at critical moments" in the whole region.

3 Topic Modelling with BERTopic

Topic modelling is a set of techniques used to automatically identify hidden thematic structures within a large collection of documents. It is a form of unsupervised machine learning that does not depend on labels or predefined categories. Over the last two decades LDA (Latent Dirichlet Allocation), first proposed by Blei *et al.* (2003), has been the

most widely used topic model for discovering latent themes in large text corpora.⁶ LDA treats each document as a bag-of-words, and by analysing word frequencies, uses a probabilistic generative model to infer which topics are likely represented in each document. However, the availability of transformer models to generate text embeddings has now made possible the use of BERTopic, a state-of-the-art topic modelling tool developed and maintained by Maarten Grootendorst (2022).⁷ BERTopic has been used to analyse a variety of corpora assembled from contemporary news sources and political discourse.⁸ Topic modelling with BERTopic relies on the same text embeddings as semantic searching. But each text is identified with a single topic (unlike LDA, which treats a document as comprised of a number of topics). Topics can be understood as vector encodings of texts that are semantically similar to each other, and that are clustered together within the high-dimensional semantic space (768 dimensions for Cohere version 2.0).

BERTopic is best described as a pipeline of text processing modules—for text embedding, dimensionality reduction, clustering, and topic representation.⁹ It takes transformer-based embeddings, applies a technique like UMAP to reduce dimensionality, clusters documents with a clustering algorithm like HDBSCAN, and then extracts representative keywords to form interpretable topics. Because

⁶ The Divergent Discourses Project (<https://research.uni-leipzig.de/diverge/>, accessed January 15, 2025) uses the iLCM (integrated Leipzig Corpus Miner) research environment, which implements topic modelling through LDA (<https://ilcm.informatik.uni-leipzig.de>, accessed January 15, 2025; see Kyogoku *et al.* 2025).

⁷ BERTopic is available as a Python library with additions and updates on github (<https://github.com/MaartenGr/BERTopic> (accessed January 15, 2025)).

⁸ Some examples: Navaretta and Hanson (2023) use BERTopic to generate topic clusters for two policy areas, Energy and Environment, in parliamentary debates and political manifestos by Danish political parties; Aenne *et al.* (2024) use BERTopic to identify dominant themes in two hashtag networks on Instagram, #blacklivesmatter and #blackouttuesday; Xing and Ni (2024) use BERTopic to investigate the portrayal of Maoism in French newspapers in the period 1963-1979.

⁹ BERTopic is named for BERT (Bidirectional Encoder Representations from Transformers) and was developed by researchers at Google (Devlin *et al.* 2019). Since 2018 the model has gone through several variations and optimisations, such as RoBERTa, and SBERT, and includes multilingual models as well.

of its modularity it allows for an enormous number of options at every step and comes with a collection of tools for visualising results. By default, BERTopic uses SBERT models from the sentence-transformers library to generate embeddings. However, other embedding models can be used, and in our case we import the Cohere version 2.0 embeddings.

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that has some advantages.¹⁰ By examining the local density of points, the algorithm automatically discovers how many clusters (topics) best describe the data, without requiring that the number of topics be specified in advance. It also does not make any assumptions about the shape of the cluster (such as assuming it is spheroid). Topics can be merged, but this is not necessarily desirable if the goal is to characterise a corpus of texts in as much detail as possible. HDBSCAN typically produces a large number of outliers (33% or more). These can be texts that are not recognisably related to any cluster, or they can be texts that fall between clusters. However, this is not a drawback in our case since our goal is not to classify every document but to identify relevant topics. BERTopic also utilises HDBSCAN to extract some “representative paragraphs” whose location within a cluster makes them the best candidates for characterising the cluster.

The clusters discovered by HDBSCAN are mathematical objects. The challenge is to assign meaningful interpretable labels to represent the topics. The default module for BERTopic at this stage in topic modelling is to apply a version of TF-IDF (Term Frequency-Inverse Document Frequency) to all of the documents within each cluster to identify words most uniquely characteristic of each topic compared to the rest of the corpus. The result is five to ten keywords which are then used to label the topic. This necessitates a word tokeniser that splits text into individual words, a list of stop words to be ignored, and a filter for punctuation. The text may also require POS tagging and lemmatisation. This is straightforward for high-resource languages

¹⁰ See <https://hdbscan.readthedocs.io/en/latest/index.html> (accessed January 15, 2025).

like English or Chinese, where many such tools exist. But without a reliable word tokeniser for modern Tibetan, it is impossible to create topic representations using TF-IDF.¹¹

However, using a generative LLM to read and summarise text and to produce representations of topics is now an option for modern Tibetan. This sidesteps some of the difficulties in developing NLP tools for modern Tibetan. After testing several LLMs, we found that Anthropic's Claude Sonnet 3.5 can read the representative paragraphs for each Tibetan topic directly and generate keywords and labels.¹² The keywords it generates come from the Tibetan source material, eliminating the need for word frequency-based keyword generation.

4 Using BERTopic with the Tibet Daily corpus

Our primary research objective is to see if this form of topic modelling will enable us to identify particular policy or political programs in the TAR, and more specifically to trace changes in their prominence over time. As human readers, we can already recognise from the repetition of certain slogans or keywords in the Chinese or Tibetan media that a policy or drive is underway, but we will not always know when such a drive began or ended. More importantly, the inevitable selectivity of natural reading can mislead us into interpreting a particular policy or drive in terms of its most prominent or striking slogan or keyword – particularly if that keyword indicates unusually repressive measures by a government – and thus overlooking other aspects of that policy or misstating its prevalence in the broader political environment.

Topic modelling with BERTopic can be used with the *Tibet Daily* corpus to offset such tendencies and misreadings. It can uncover the major themes or topics that underly collections of documents,

¹¹ The Divergent Discourses project is developing a Tibetan language model for spaCy that includes a vocabulary tokeniser and POS tagging capabilities to process Tibetan text for input into iCLM. See Kyogoku *et al.* 2025.

¹² <https://www.anthropic.com/claude/sonnet> (accessed January 15, 2025). Claude Sonnet is accessed through an API and accepts detailed instructions through a prompt incorporated into the code.

presenting aspects or purposes of a policy that might not be evident to a casual reader, allowing more precise and comprehensive forms of discourse analysis. The occurrence of topics can be modelled dynamically, revealing when they become prominent and when discussion in the official media tapers off, indicating the life-cycle of each policy or drive. Since topic modelling is based on the same technology as semantic searching, it can identify texts that address a given policy even though none of the keywords normally associated with that policy are used – an important tool for an analyst, since policies and official efforts at ideological promotion in China (and Tibet) are usually put into practice some time before officials settle upon their slogans or keywords. In addition, topic modelling of this type can display the relative prominence or otherwise of a policy within the larger policy environment, revealing themes that might be striking to a casual reader but are perhaps relatively infrequent in overall discourse, or the opposite.

Applying topic modelling to an undefined set of texts, such as the entire contents of a newspaper over a given period, is generally not effective, because the tool will tend to return results that are so general as to be already obvious to the reader, such as “news”, “sports”, “arts” and so forth. We, therefore, selected a particular question of interest and applied the tool only to texts relevant to that question. In this case, we selected “religion” as our overarching question. Our purpose was to see if the tool would provide more insights about the Chinese government’s policies towards religion in the TAR than we had already gathered from unstructured readings of articles over recent years. Those readings had led us to note already a number of key terms or slogans found frequently in official articles and speeches about religion. One example of a religion-related policy term is the phrase “four standards” (ཚད་གཞི་བཞི), discussed earlier in relation to semantic searching, which had seemed to us particularly prominent in our unstructured readings. This term refers to a set of behaviours or attitudes required of all monks and nuns in the TAR. These require the monks to be “politically reliable”, “accomplished in religious knowledge”, “convincing in morality”, and to “play an active role at critical moments” (Chang & Chen 2020). We knew already that these

requirements had been introduced in about 2016 or shortly after (HRW 2018), but we were unsure if the policy would remain in force eight years later. If so, that would be unusual because CCP drives or policies often disappear from public view within two or three years. More importantly, the keywords or formulations (提法 *tifa*) used for each policy or drive are almost never explained in public documents. In the case of the “four standards”, two appear to be about encouraging religious knowledge and ethical conduct, but other religious policies, as we shall see, imply stringent limitations on the actual meanings of these terms. The definition of “an active role at critical moments” could refer to denouncing any others who have dissident opinions or could refer to obeying official orders at the time of the Dalai Lama’s death; it has never been publicly explained. We hoped that topic modelling, combined with semantic searching, would increase our chances of learning more about this drive and its relation to similar policies at the time.

Our first step was to create a subcorpus with a manageable number of articles. To achieve this, we collected all the articles from *Tibet Daily* between 2014 and 2024 in which terms for “religion” appear. We limited ourselves in this study to using the Tibetan-language corpus of *Tibet Daily* rather than the Chinese-language corpus so as to demonstrate how source material in modern Tibetan can be effectively explored using this set of tools.

The articles have been split into numbered paragraphs and then, where necessary, split into chunks of 256 tokens or less. These were filtered paragraph by paragraph for entries that satisfied the Boolean search query: (“བོད་བརྗེད་ནང་བསྟན” OR “ཚོས་ལུགས”) — “Tibetan Buddhism” OR “Religion”. Filtering for these two strings (which allows for their appearance in longer phrases) effectively captured all of the paragraphs in which religion was mentioned in some context. The filtered results were then arranged into a CSV file for each year. In total, the religion subcorpus comprised 3,952 articles with 6,622 paragraphs over the 2014–24 period (divided into 9,048 chunks that do not exceed 256 tokens).

In Figure 1 we show a chart, produced using a keyword (string) search on the CSV files in our subcorpus, that displays the relative

frequency of articles (with one or more paragraphs) that satisfy the Boolean search query by year. Articles including one or other of the two terms we used for religion comprise roughly 5% to 7% of all articles between 2014 and 2024. The chart shows, however, a dramatic drop off in the number of such articles from 2022 onwards (as low as 3.5%), an outcome of which we had been completely unaware despite regular reading of the Chinese media in Tibet.

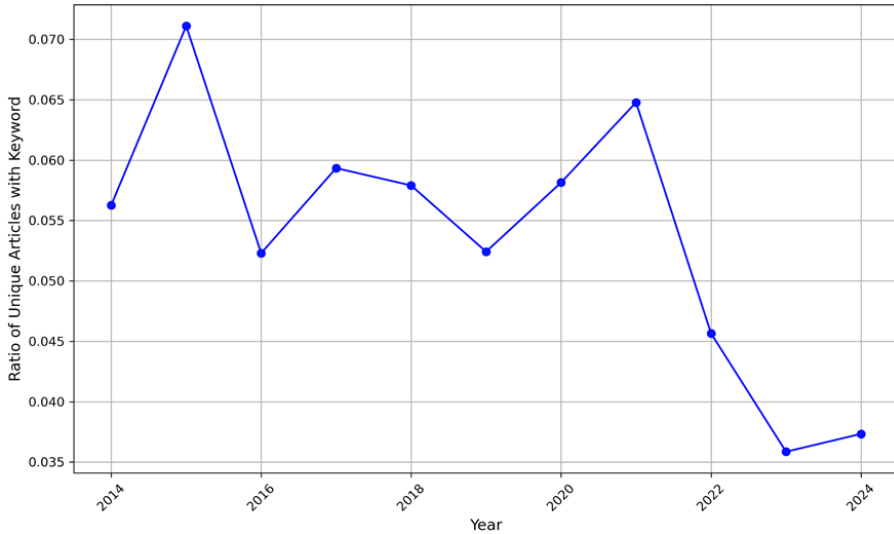


Figure 1 Relative frequency of articles containing the terms བོད་བརྗེས་ནང་བཟུང་ (Tibetan Buddhism) or ཚོས་ལུགས་ (religion) in Tibet Daily, 2014–24

We then imported this filtered subcorpus along with the aligned Cohere version 2.0 embeddings for each paragraph chunk into BERTopic. This resulted in 121 topics (0 to 120). There were 4,368 outlier chunks (48.3% of the total of 9,048 chunks) which BERTopic categorised as outliers (it assigns the topic number -1 to them). BERTopic produced what it identified as the ten most representative paragraphs for each topic, and we then used Claude Sonnet 3.5 to read each of these sets of ten paragraphs and to generate topic labels and keywords for them. This produced a label and five keywords in Tibetan, together with English translations, in order of descending importance for each topic. The keywords it selected are not necessarily found in all ten of the representative paragraphs, but taken together

these keywords provide a semantic representation of the ten clustered representative paragraphs for each topic. Table 2 displays a list of the top 21 topics (topics #0 to #20) in descending order of frequency along with the keywords and labels produced by Claude Sonnet 3.5.

Table 2: Topics #0 to #20 with Keywords and Labels

Topic	Count	Keywords	Label
0	256	Religious freedom policy (ཚོས་དང་རང་མོས་སྲིད་ཇུས་), monastery management (དགོན་སྡེ་འདོད་དམ), ethnic unity (མི་རིགས་མཐུན་སྦྲེལ་), legal supervision (ཁྲིམས་ལྟར་དོད་དམ), religious affairs work (ཚོས་ལུགས་ལས་དོན)	Implementation of party's religious policy and monastery management
1	133	Ethnic unity (མི་རིགས་མཐུན་སྦྲེལ་), environmental protection (སློང་བཟུང་ཁོར་ཕྱག), religious harmony (ཚོས་ལུགས་ཞི་མཐུན), social stability (སྤྱི་ཚོགས་བརྟན་ལྗིང), people's livelihood (དམངས་འཚོ)	Social harmony and development in Tibet Autonomous Region
2	124	Four standards (ཚད་གཞི་བཞི), model monks and nuns (སྦྲོན་ཐོན་གྲྭ་བཙུན་མོ), religious affairs regulations (ཚོས་ལུགས་ལས་དོན་གྱི་སྲོལ་ཡིག), socialist adaptation (སྤྱི་ཚོགས་རིང་ལུགས་དང་འཚམ་མཐུན), Xi Jinping thought (ཞི་ཅིན་ཕིང་གི་དགོངས་པ)	Implementation of four standards policy in Tibetan Buddhist monasteries
3	121	Reincarnation system (སྐུལ་སྐྱེའི་ཡང་སྲིད), government approval (ཡུང་དབྱང་སྲིད་གཞུང་གི་ཚོག་མཚན), legal regulation (ཁྲིམས་སྲོང), traditional procedures (ཚོས་ལུགས་ཀྱི་ཚོག), domestic search requirement (རྒྱལ་ནང་ནས་ཡང་སྲིད་ཚུལ་འཚོལ་བ)	Chinese government control over Tibetan Buddhist reincarnation system
4	113	Religious harmony (ཚོས་ལུགས་འཚམ་མཐུན), monastery administration (དགོན་སྡེ་འདོད་དམ), monk welfare (གྲྭ་བཙུན་གྱི་འཚོ་བ), standardised management (ཚད་ལྡན་དོད་དམ), social insurance (འགན་བཅོལ)	Monastery management and religious harmony implementation policies

Topic	Count	Keywords	Label
5	113	United front work (འཐབ་ཕྱོགས་གཅིག་གྱུར་), religious affairs committee (ཚོས་ལུགས་ལས་དོན་ལྷན་ཁྲུང་), reincarnation training (སྐུལ་སྐྱེའི་གསོ་སྦྱང་), monastery management (དགོན་སྡེའི་དོ་དམས་), Buddhist interpretation (ནང་བསྟན་གྱི་དགོངས་པ་གསར་འགྲེལ་)	Religious and ethnic affairs meetings and training in tibet Autonomous Region
6	104	Patriotic religious devotion (རྒྱལ་གཅིས་ཚོས་གཅིས་), social harmony (སྤྱི་ཚོགས་ཞི་མཐུན་), monastic discipline (སྤྱི་གཞན་ལམ་སྲུང་སྦྱོང་), ethnic unity (མི་རིགས་མཐུན་སྦྲིལ་), Buddhist traditions (བོད་བརྒྱུད་ནང་བསྟན་གྱི་སྲིལ་རྒྱན་)	Buddhist monastics' patriotic and religious development in Tibet
7	100	Social stability (སྤྱི་ཚོགས་བརྟན་ལྷིང་), religious affairs management (ཚོས་ལུགས་ལས་དོན་དོ་དམས་), border security (མཐའ་མཚམས་བདེ་འཇགས་), anti-separatism (ལ་ཕྲལ་ལ་དོ་ཚོལ་), public safety (སྤྱི་ཚོགས་བདེ་འཇགས་)	Social security and religious affairs management in Tibet
8	98	Separatist politics (ལ་ཕྲལ་རིང་ལུགས་), Tibetan independence (བོད་རང་བཙོན་), religious exploitation (ཚོས་ལུགས་ཀྱི་ཕྱི་གོས་), social disruption (སྤྱི་ཚོགས་ཟེར་ཟིང་), anti-China forces (གྲུང་གོར་དོ་ཚོལ་)	Chinese government criticism of 14 th Dalai Lama's political activities
9	95	Economic development (དཔལ་འཕྲོར་འཕེལ་རྒྱས་), ethnic unity (མི་རིགས་མཐུན་སྦྲིལ་), religious harmony (ཚོས་ལུགས་འཆམ་མཐུན་), social stability (སྤྱི་ཚོགས་བརྟན་ལྷིང་), ecological protection (སྐྱེ་བསམས་སྲུང་སྦྱོང་)	Tibet's modern development and social harmony progress report
10	91	Party gratitude (ཉང་གི་བཀའ་རྒྱུན་), rational religious understanding (ཚོས་ལུགས་ལ་དཔྱད་ཤེས་), poverty alleviation (དབྱེ་སྦྱོང་), happy life (བདེ་སྤྱིད་འཚོ་བ་), ethnic unity (མི་རིགས་མཐུན་སྦྲིལ་)	Party loyalty and religious moderation in economic development
11	86	<i>Thangka</i> paintings (ཐང་གཤམ་), religious artistry (ཚོས་ལུགས་སྐྱུ་རྩལ་), traditional techniques (སྲིལ་རྒྱན་ལག་རྩལ་), monastery displays (དགོན་པར་བཤམས་), artistic periods (དུས་མཚམས་)	Historical development and artistic traditions of Tibetan <i>thangka</i> painting

Topic	Count	Keywords	Label
12	86	Radio translation (རྒྱ་རྒྱུ་ལྷན་ཁྲིམས་ལྷན་ཁྲིམས་), cultural adaptation (རིག་གནས་སྲིལ་རེས་), news media (གསར་འགྲུར་བརྒྱུད་ལས་), target audience (གསན་པ་ལོ་), translation accuracy (ཡང་དག་པའི་སྒྲུང་ཐབས་)	Translation principles and cultural exchange in Tibetan media broadcasting
13	85	Religious criticism (ཚོས་ལུགས་ལ་དཔྱད་ཤེས་), negative influence (གྲགས་ཀྱིན་རན་པ་), present happiness (དེ་ཆའི་བདེ་སྲིད་), show the flag [one's political stance] (དར་ཆ་གསལ་སྟོན་), socialist adaptation (སྤྱི་ཚོགས་རིང་ལུགས་དང་འཛམ་མཐུན་)	Countering religious influence of 14 th Dalai Lama through socialist education
14	85	Religious freedom (ཚོས་དད་རང་མོས་), legal religious activities (ཁྲིམས་མཐུན་ཚོས་ལུགས་བྱེད་སྟོན་), counter terrorism (འཛིགས་སྐྱུལ་རིང་ལུགས་ལ་རོ་ཤོལ་), constitutional compliance (བཅའ་ཁྲིམས་སྲུང་བཅི), religious harmony (ཚོས་ལུགས་འཆམ་མཐུན་)	Religious activities regulation and anti-terrorism legal framework
15	83	Religious freedom (ཚོས་དད་རང་མོས་), monastery supervision (དགོན་སྡེ་དོ་དམ་), patriotic religion (རྒྱལ་གཅེས་ཚོས་གཅེས་), social harmony (འཆམ་མཐུན་བརྟན་སྲིད་), Buddhist education (ནང་བསྟན་སྲོལ་གྲིང་)	Buddhist monastery management and religious policy implementation in Tibet
16	74	Socialist adaptation (སྤྱི་ཚོགས་རིང་ལུགས་དང་འཛམ་མཐུན་), religious harmony (ཚོས་ལུགས་འཆམ་མཐུན་), monastic management (དགོན་སྡེ་དོ་དམ་), Tibetan studies (བོད་རིག་པ་ཞིབ་འཇུག་), academic research (ཐོས་བསམ་ཞིབ་འཇུག་)	Adaptation of Tibetan Buddhism to socialist society and academic development
17	72	Patriotic Buddhism (རྒྱལ་གཅེས་ཚོས་གཅེས་), religious harmony (ཚོས་ལུགས་འཐུན་སྲིལ་), national unity (མེས་རྒྱལ་གཅིག་ཁྱུར་), Buddhist traditions (བོད་བརྒྱུད་ནང་བསྟན་), social development (སྤྱི་ཚོགས་འཕེལ་རྒྱས་)	Panchen Lama's role in Tibetan Buddhism and Chinese socialist society
18	68	Religious patriotism (རྒྱལ་གཅེས་ཚོས་གཅེས་), social harmony (ཞི་མཐུན་སྤྱི་བཀའ་), Buddhist reform (ཚོས་ལུགས་སྒྱུར་བཅོས་), monastic discipline (སྡེ་ཁྲིམས་), social development (སྤྱི་ཚོགས་འཕེལ་རྒྱས་)	Adapting Tibetan Buddhism to modern socialist society

Topic	Count	Keywords	Label
19	64	Rational religious understanding (ཚམས་ལུགས་ལ་དཔྱོད་ཤེས་), happy life (བདེ་སྲིད་འཚོ་བ་), educational guidance (སློབ་གསོ་ཇིང་སྟོན་), hard work and effort (དཀའ་ལྷན་འབད་འཐབ་), reducing religious superstition (ཚམས་ལུགས་ཀྱི་ཕན་མེད་ལུགས་ཀྱི་སེལ་)	Religious education and rational approach to modern life
20	59	Monastery supervision (དགོན་ཕྱེད་དཔལ་), religious harmony (ཚམས་ལུགས་འཆམ་མཐུན་), monastic welfare policies (དགོན་ཕན་གྱ་ཕན་སྲིད་ཅུས་), monks and nuns (གྲུ་བཟུན་), religious development (བསྐྱེད་དོན་ལམ་སྟོར་)	Monastery management and religious policy implementation in Tibet

The information in Table 2 allows us to draw up an overview of religious policy in the TAR during the study period. It shows that “four standards” were indeed an important part of the religious policy environment – they appear as Topic #2 in the clustering performed by BERTopic. In addition, the labels and keywords chosen by Claude Sonnet 3.5 confirm that the four standards are directed at monks and nuns (usually referred to as “religious professionals” in Chinese legal documents) and are part of a regulatory program (“Religious Affairs Regulations”). The inclusion of the word “implementation” in the label given to this topic strongly indicates that the four standards drive involves active engagement by officials in monasteries and nunneries and is intended to require compliance of some sort from its targets immediately (presumably under threat of some kind of institutional punishment, such as expulsion from a monastery). If we look for other instances of implementation-oriented or regulatory drives among these 21 topics, we can see that the labels and keywords for four other topics – #0, #4, #15 and #20 – include the word implementation. This indicates that these topics too primarily concern the management of monasteries (or nunneries) and the imposition of regulations on their personnel. Five other topics (#3, #5, #7, #14 and #16) refer to regulations or management.

Using the labels generated by Claude Sonnet 3.5, we classified all of the 121 topics as belonging to one of four broad categories indicating their political or discursive purpose:

- (1) regulatory topics relating to management or administration;
- (2) celebratory topics praising cultural and social achievements of the state without necessarily declaring a political message;
- (3) ideological-positive topics presenting aspirational concepts, goals and principles; and
- (4) ideological-negative topics attacking or critiquing prevailing beliefs and practices that are to be opposed and eradicated.

In total, regulatory topics comprised 49.9% of paragraph chunks assigned to a topic; celebratory topics represented 8.7% of paragraph chunks; 31.9% belonged to ideological-positive topics; and 9.5% were categorised as ideological-negative topics.

This allows us to hypothesise, if official media pronouncements are indicative, that about one half of religious policy is focused on direct intervention by officials in monastic life and practice or on monastic procedures such as the identification and recognition of reincarnated lamas. These policy drives are thus explicitly directed at strengthening management by the state of “monastic professionals” and their institutions. In terms of print space (or numbers of chunks) dedicated to them in *Tibet Daily*, six of the top eight topics identified by BERTopic involve the implementation of regulations of this kind, primarily on monasteries.

This recalls Pitman Potter’s observation (2003) that religious policy in China in the post-Mao era shifted in the 1990s to a focus primarily on the “management” of “religious personnel” and institutions. Our topic model shows that a focus on the management and regulation of monks and nuns has remained a prominent part of religious policy in the TAR. However, we can also see a significant difference between the current and the earlier regulatory regimes. In the 1990s (specifically from 1996 onwards), the requirements imposed on Tibetan monks and nuns in the TAR required them primarily to supply memorised responses to a written examination, culminating with a declaration of patriotism and a formulaic denunciation of the Dalai Lama (Barnett & Spiegel 1996).

What constituted compliance was thus relatively clear. The four standards (“political reliability”, “accomplishment in religious

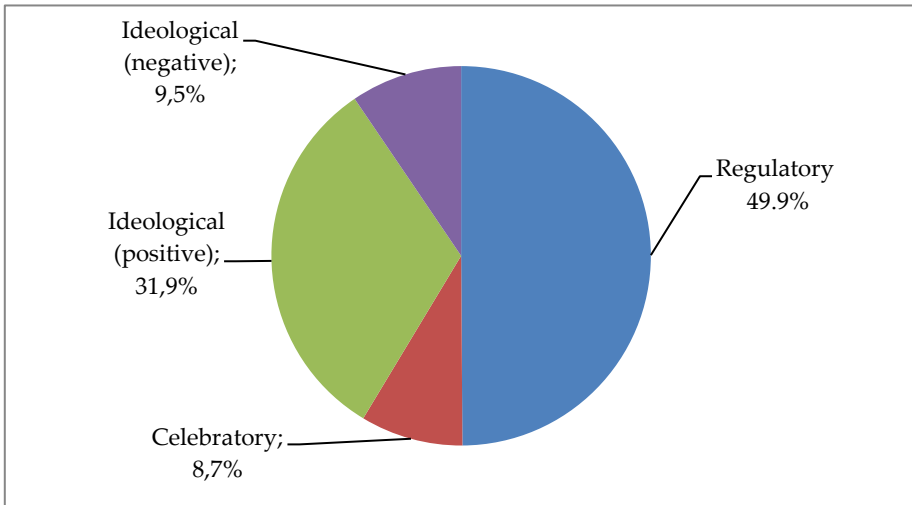


Figure 2 Proportion of paragraph chunks in the Tibet Daily containing regulatory, celebratory or ideological topics relating to religion, 2014–24.

knowledge”, “convincingness in moral conduct”, and being “active at critical moments”) are, however, different in that respect from the 1990s requirements: it is doubtful that they can be legally defined, and compliance would seem hard to measure. There are no signs that formulaic recitation, such as in an examination, is accepted by officials in the current drives as proof of adherence to official demands. Instead, the four standards appear to be a new form of regulatory practice that requires changes to thought, including to thought in the future (“critical moments”), as much as to present behaviour. This reflects the overwhelming shift in Tibet policy, and in policy regarding other “minorities”, under the leadership of Xi Jinping to prioritising the management of thought, not just behaviour (this principle is referred to in some CCP documents as “correctly handling the relationship between ‘controlling the stomach’ and ‘controlling the brain’”; see Chang & Chen 2020).

What, then, does the topic model suggest about the other aspects of religious policy in the region over the last decade? The model does show that some aspects of religious policy are what we might call celebratory, such as promoting “translation accuracy” in media reports (#12), and promoting traditional religious artwork and

painting (#11). These, however, are relatively minor topics, covering only some 8.7% of the text (chunks) in the study corpus. As for the remaining forms of religious policy identified by BERTopic, totalling 41.4% of the paragraph chunks assigned to topics, we can see that they are not regulatory – they do not focus primarily on the imposition of regulations and do not single out monastic institutions as their targets. Instead, they are focused on the promotion of certain concepts, such as “ethnic unity” (མི་རིགས་མཐུན་སྦྲེལ་), “religious harmony” (ཚོས་ལུགས་ཞི་མཐུན་), and “social stability” (སྤྱི་ཚོགས་བརྟན་ལྗིད་). These concepts are not linked in the labels or keywords to regulations or to monastic management, so we can infer that these policies consist of ideological rather than regulatory drives. That is, they are efforts by the state to inculcate, through some sort of educational process, these concepts in the minds of the target group. That group is not primarily “religious professionals”, but the general population. In particular, we can see from the inclusion of such concepts as “people’s livelihood” (དམངས་འཚོ་), “happy life” (བདེ་སྲིད་འཚོ་བ་), and “hard work and effort” (དཀའ་སྤྱད་འབད་འཐབ་), that these ideological or educational drives are generally aimed at lay believers. If we look closer, we can also see that these drives are not colour-blind – in several cases, such as those about reincarnation, the Dalai Lama or the Panchen Lama, these drives are specifically targeting only Tibetan Buddhists, not followers of other religions. As far as we can tell, these drives may be exhortatory rather than disciplinary – that is, they may not involve explicit punishments or threats (or at least not legal ones).

This emphasis on mass inculcation of certain concepts reflects the instructions given by Xi Jinping at the Seventh Central Forum on Tibet Work in August 2020, where he defined China’s overall Tibet policy for the coming decade:

Xi Jinping pointed out that Tibet work must adhere to the focus and focus on maintaining the unity of the motherland and strengthening ethnic unity. We must strengthen education and guidance for the masses, widely mobilise the masses to participate in the anti-secession struggle, and form an iron wall to maintain stability. We must carry out in-depth education on the history of the Party, the history of New China, the history of reform and opening up, and the history of socialist development, and carry out in-depth education on the history of the

relationship between Tibet and the motherland, and guide the people of all ethnic groups to establish a correct view of the country, history, nation, culture, and religion. We must attach importance to strengthening ideological and political education in schools, run the spirit of patriotism throughout the entire process of education at all levels and types of schools, and plant the seeds of love for China in the hearts of every young person. We must cultivate and practice the core socialist values, and constantly enhance the identification of the people of all ethnic groups with the great motherland, the Chinese nation, Chinese culture, the Communist Party of China, and Socialism with Chinese characteristics. ...We must actively guide Tibetan Buddhism to adapt to socialist society and promote the sinicisation of Tibetan Buddhism. (Xinhuanet 2020)

Before Xi, Tibet policy had generally targeted selected sub-groups of the Tibetan population (primarily monks and nuns, returnees from exile and certain types of intellectuals) as suspected dissidents and subjected them to control and re-education; under Xi, from the Seventh Forum onwards, the primary focus became the “education and guidance” of the Tibetan population as a whole. The topic model shows that a third of the texts in the *Tibet Daily* subcorpus described drives pursuing this new priority.

The labels and keywords generated by Claude Sonnet 3.5 for each topic point to an important distinction among these mass ideological drives: some of them involve positive incentives, while others are negative. Negative propaganda or indoctrination – explicit attacks on or critiques of religious belief – have long been viewed within the CCP as a high-risk strategy when it involves religion. That was the principal reason for Mao’s conciliatory approach to central Tibetans in the 1950s, and for “Document 19”, the famous reformist statement on religion issued by the CPP in 1982, which condemned the anti-religion policies of the Cultural Revolution and called for an end to any attempts by the state to eliminate religion (MacInnis 1989).

Our topic model shows, however, that negative propaganda about religion has become a significant part of the current religious policy environment, although it is confined – at least in print – to a secondary role. Of the texts (chunks) identified by BERTtopic as parts of

ideological drives, 31.9% advance positive goals or concepts, such as “ethnic unity”, “religious harmony”, “social stability”, “patriotic devotion”, “gratitude to the Party”, and “prosperity”. By contrast, the remaining 9.5% appear to signal attacks on certain forms of belief or practice. They include references to “rational religious understanding” (ཚོས་ལུགས་ལ་དཔྱད་ཤེས) and the “negative influence [of religion]” (གུགས་རྒྱན་རན་པ) in topic #13, and to “reducing religious superstition” (ཚོས་ལུགས་ཀྱི་ཕན་མེད་གུགས་རྒྱན་སེལ) in topic #19. Similarly, topic #8 appears to be a condemnation of some form of religious view – the label for this topic indicates that this view is related to support for the Dalai Lama – as “religious exploitation” (ཚོས་ལུགས་ཀྱི་ཕྱི་གོས) because it is linked to “separatist politics” (ཁ་ཕྱལ་རིང་ལུགས), “Tibetan independence” (བོད་རང་བཙན), “social disruption” (སྤྱི་ཚོགས་ཟར་བྱིང), “anti-China forces” (ཡུང་གོ་རྩོམ་པ་).

In addition, references in the official media to apparently positive concepts such as “rational religious understanding” (ཚོས་ལུགས་ལ་དཔྱད་ཤེས), “happy life” (བདེ་སྤྱིད་འཚོ་བ), “hard work and effort” (དཀའ་སྤུང་འབད་འཐབ), and “present happiness” (དེ་སྐབས་བདེ་སྤྱིད) are in practice negative critiques of religion: we know from readings of articles on these topics that “rational religious understanding” and related terms are key parts of critiques of any religious practices deemed excessive, such as offerings to religious figures or institutions. Similarly, “present happiness” is a reference to the Party’s current drive to persuade Tibetans that lay religious belief should never include considerations of one’s future after death. The underlying negative context of the “present happiness” concept is shown by the label for this topic (#13), which describes it as “Countering Religious Influence of 14th Dalai Lama”. The topic model thus indicates efforts to present negative critiques of religion in positive terms, but also a failure so far by officials in at least some drives to avoid direct attacks on religious behaviour.

5 *Dynamic Topic Modelling*

BERTopic also provides a collection of tools for analysing the evolutions of topics over time through what is known as dynamic topic modelling. Dynamic topic modelling looks at the distribution of the

documents (paragraph chunks) in a topic cluster over a number of timesteps. Spikes in frequency for a topic can signal the occurrence of a political campaign or a new emphasis on propaganda. The keyword representations for topics can be read as the discursive elements that define these developments. Using dynamic topic modelling, BERTopic distributed each of the 121 topics into 11 bins for the period from 1 January 2014 to 14 November 2024. Here we will discuss some of the important topics selected from the first 21 topics (#0 to #20).

Figure 3, for example, which plots topics #13 and #19, allows us to establish a timeline for negative campaigning by officials with regard to religious belief among lay Tibetans. It also shows the discursive interplay of positive and negative language in campaigns. The focus of topic #13 (“Countering Religious Influence of 14th Dalai Lama Through Socialist Education”) is the “negative influence” of an allegedly backward version of Tibetan Buddhism; but it tries to express this critique in a positive way. It does this by stressing the need for a “rational understanding” of religion that will produce happiness in one’s current life. References to this topic or drive first appeared in *Tibet Daily* in 2018 and peaked in 2020 at the time of Xi’s address to the Seventh Forum on Tibet Work. References to this theme fall off quickly by the end of 2022. This suggests that policies are often introduced in a local region some two years or more before a central leader announces them to the public; a meeting such as the Seventh Forum is thus in many ways a confirmation of already existing policies. We see here that an ideological drive or campaign of this sort is relatively short, in this case lasting for around two years, suggesting that they are a response to an immediate but likely short-term instruction from the leadership.

Figure 3 also shows that topic #19 (“Religious Education and Rational Approach to Modern Life”) is coterminous with topic #13. Topic #19 is directed to cadres, instructing them to use a positive form of ideological education regarding religious belief by educating the masses to devote themselves to hard work in this life in order to achieve happiness. Its aim, however, is again a negative one, the “reduction of religious superstition”. These two drives, both active between 2018 and 2022, illustrate the use of both positive and negative

discursive forms in propaganda, but it is clear that the anti-Dalai Lama drive at this time was significantly more prominent than the drive to persuade believers to focus exclusively on their “present happiness”.

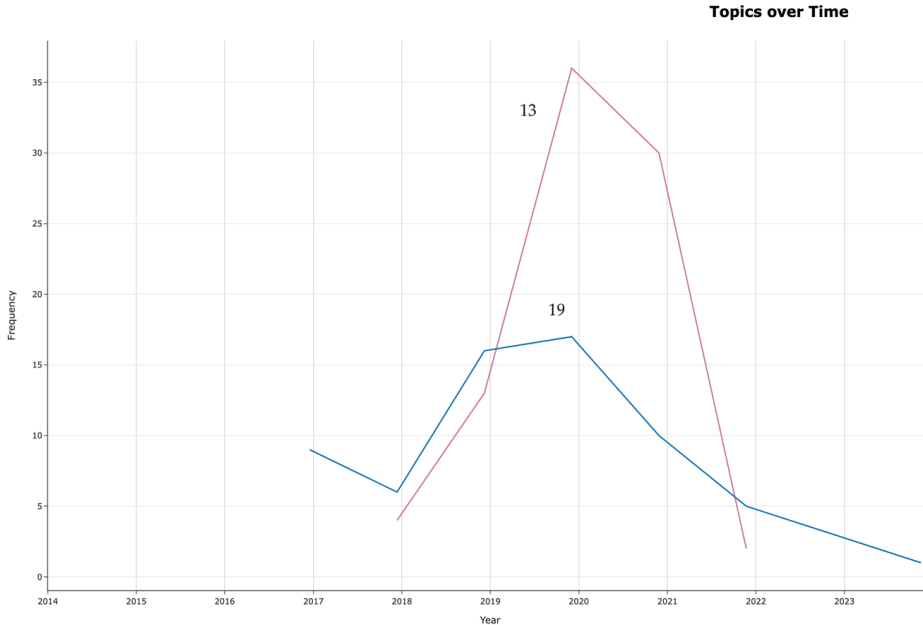


Figure 3 Frequency in the Tibet Daily of topics #13 (“Countering the 14th Dalai Lama”) and #19 (“Religious Education and Rational Approach to Modern Life”), 2014–24.

Figure 4 provides timelines for three drives that had somewhat longer lifespans. These were either regulatory drives or were fundamental to China’s long-term Tibet policy. The first and most prominent is the drive to impose governmental regulations concerning the selection of reincarnate lamas (topic #3). This became a major political priority for the Chinese state after the exiled Dalai Lama unilaterally declared his recognition of a child as the 11th Panchen Lama in May 1995; this led China to impose formal regulations in 2007 abrogating to itself alone the right to choose or appoint reincarnate lamas. The drive to enforce compliance with these regulations accelerated after 2011, when China appears to have begun preparing for the death of the current Dalai Lama. We see accordingly that this drive was already in process at the start of the period covered by our subcorpus in 2014, that it peaked in 2019 and again to a lesser extent in 2021, and still continues. The long-

running nature of this drive is as expected, but the peak in 2019 has yet to be explained, and, once again, the marked decline in visibility of this drive (and of any *Tibet Daily* articles referring to religion) after 2022 is surprising.

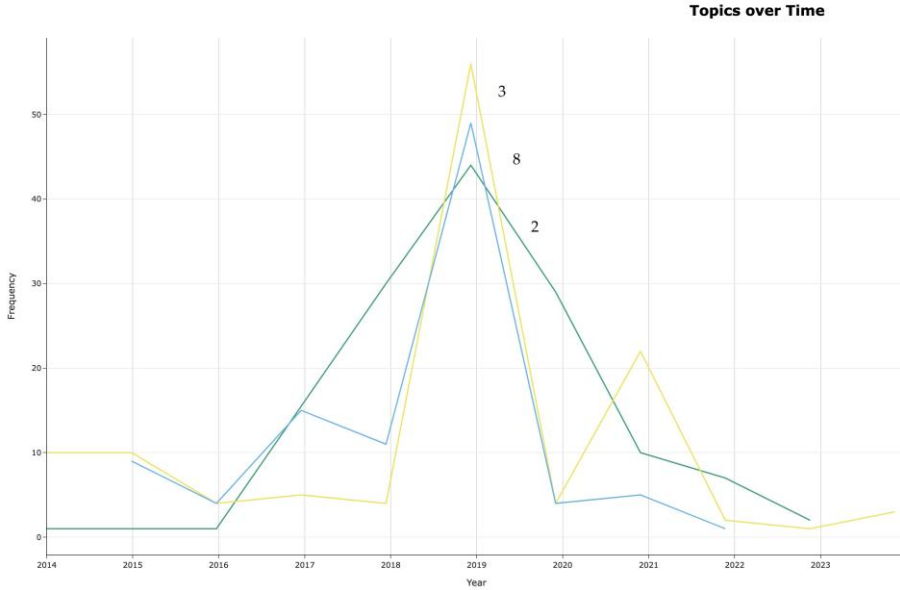


Figure 4 Frequency in the *Tibet Daily* of topic #3 (“Government Control Over Tibetan Buddhist Reincarnation System”); topic #8 (“Criticism of the 14th Dalai Lama’s Political Activities”); and topic #2 (“Implementation of Four Standards Policy in Tibetan Buddhism”), 2014–24.

The second timeline shown in Figure 4 maps a closely related topic, the drive to denounce “the 14th Dalai Lama’s Political Activities” (topic #8). This, the most striking and controversial of all Chinese policies in Tibet, was initiated at the Third National Forum on Tibet Work in July 1994, is generally seen as in many ways the bedrock of China’s political strategy in Tibet. We see here that it continued into the 2020s. Surprisingly, however, it seems to have been out of sight in 2014, to have re-emerged in 2015, and to have peaked in 2019 shortly before the Seventh Forum, exactly at the same time as the drive to promote the reincarnation regulations. Denunciations of the Dalai Lama (or “the Dalai”, as he is referred to in the official Chinese media) in *Tibet Daily*, at least in the form identified by BERTopic here, then

disappeared from view shortly afterwards, with no references since 2022. These again are findings that have not previously been noted.

The third topic shown in this figure is the four standards drive (topic #2), which, as we have seen, is a regulatory drive that imposed new requirements on monks and nuns. The timeline shows that it was already appearing as a topic in *Tibet Daily* by 2014, at least two years before the time when foreign observers had believed it to have begun. References to this drive do not appear after 2023. Though these three topics are directed at different targets (the “four standards” campaign in topic #2 is for the monasteries, while the anti-Dalai Lama drive and the promotion of reincarnation regulations are society-wide), they overlap very precisely, all peaking at the same time in 2019. We can see that the recent peak in these regulatory or long-running drives (the anti-Dalai Lama and the reincarnation drives would certainly have shown earlier peaks in the period before 2014, if our subcorpus had included those years) occurred a year before the Seventh Forum and the corresponding peak in the ideological drives that we saw in Figure 3. Overall, we can see that drives run in tandem: at a time when the CCP activates a push on religion in the TAR, that push will consist of multiple components and subsidiary drives more or less simultaneously.

Dynamic topic modelling also shows topics which are not marked by a single peak or a short duration, but are long-standing and recurring. These are distributed more evenly over a number of years and indicate a discourse that is sustained over a longer period or repeated regularly. Some examples are displayed in Figure 5; all these topics or drives continued throughout our research period and were still being referred to in *Tibet Daily* as of 2024. The most prominent concerns monastery management (topic #0). This is the highest frequency topic identified by BERTopic.. It is a regulatory drive directed at monasteries, which, while restating the principle of respecting religious beliefs and allowing normal religious activities, cracks down on illegal activities and calls on cadres to implement and strengthen the Party’s management of religious institutions. It is present throughout the entire period from 2014 to 2024, rising from 2015 onwards, falling off to some extent by 2020, with a sustained peak

between 2016 and 2018, and again in 2022. It exemplifies the principle that we have already seen that religious policy in Tibet is at its basis the persistent imposition of managerial control over monks and nuns. When the monastery management drive drops off slightly in 2019, we can see that it is replaced by the peak in the four standards drive (topic #2), which is only a new and more demanding form of monastic management, as well as by the drives requiring compliance with the anti-Dalai Lama drive and the reincarnation regulations.

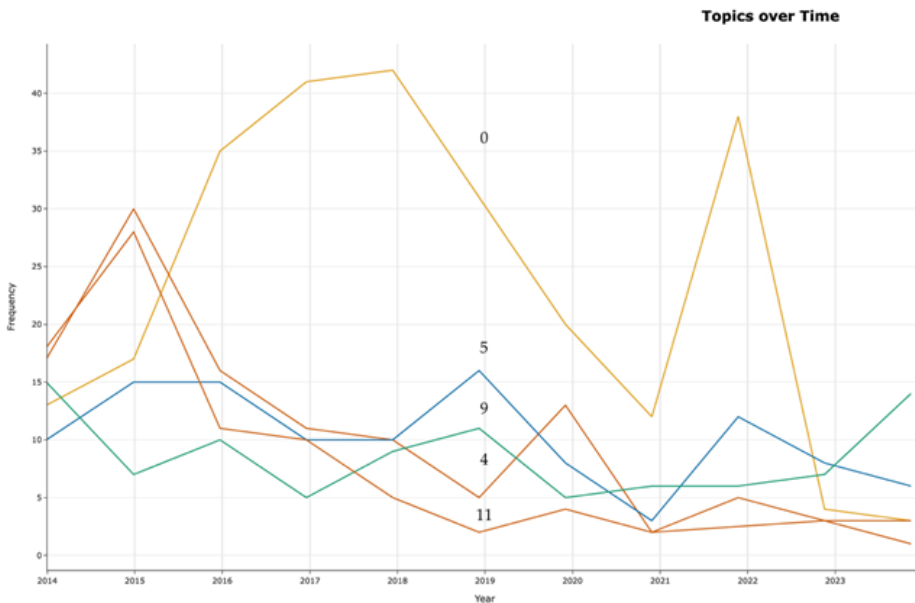


Figure 5 Frequency in the Tibet Daily of topic #0 (“Implementation of Party’s Religious Policy and Monastery Management”); topic #4 (“Monastery Management and Religious Harmony Implementation Policies”); topic #11 (“Historical Development and Artistic Traditions of Tibetan Thangka Painting”); topic #5 (“Religious and Ethnic Affairs Meetings and Training in Tibet Autonomous Region”); and topic #9 (“Tibet’s Modern Development and Social Harmony Progress Report”), 2014–24.

The timelines in Figure 5 show a number of other topics or drives that are long-running and persistent, marked by continuity rather than peaks. These topics include two subsidiary components of the monastery management drive that evidently are more or less in continuous operation. One – “Monastery Management and Religious Harmony Implementation Policies” (topic #4) – is basically a reminder to cadres to present, as its label indicates, a positive aspect of monastic

regulation by emphasising the state's provisions of health and welfare support for (officially recognised) monks and nuns. The other is the ongoing training and oversight of the cadres who implement religious affairs policies in Tibet, under the leadership of the Party agency known as the United Front Work Department (topic #5).

Other "persistent" topics shown in Figure 5 are one celebrating Tibetan religious painting (topic #11), and one linking "religious harmony" to economic development (topic #9). These topics are more likely to represent consistent themes to be maintained in propaganda work rather than drives. Here again we see an effort to emphasise positive forms of propaganda and management.

6 Conclusion

This discussion of topic modelling has focused on just a few of the higher frequency topics that were identified in our research corpus. But even from this selection, we can reach a number of provisional hypotheses. One is that high-frequency topics in some cases will represent political drives or campaigns where officials are mobilised to achieve a specific outcome in one or other sector of society. Such drives will typically be of relatively short duration, perhaps of one to two years in some cases, and will be marked by peaks in terms of frequency of references in the media. These will often be responses to instructions or calls from a central leader, and may be intended to signal highly visible compliance by local officials to national-level instructions.

In general, we noted two kinds of drives of that type: regulatory ones aimed mainly at religious institutions and professionals, and ideological ones that aim to change thought and attitude among the wider public. The regulatory drives, and above all monastery management, appear to be the basic, ongoing or staple element of religious policy in the TAR. The regulatory drives will often be of relatively long duration, and will include multiple subsidiary drives.

The topics that indicate ideological drives, designed to inculcate a particular concept or opinion among the population, will typically be

shorter in duration and stronger in intensity. These drives appear to be a dominant feature in the Xi Jinping era, since the entirety of minority populations are now deemed in need of radical political re-education and improvement. They show recurrent attempts to present negative critiques of religion in positive terms.

Topics which involve denunciations of the Dalai Lama or of Tibetan independence (and required compliance with reincarnation regulations) are an exception to the principle of avoiding negativity in religious discourses. These topics show peaks of activity but are persistent over time. This appears to reflect the persistent perception among officials of the Dalai Lama and the independence concept as the core threat to China in Tibet, the core assumption on which all Tibet policy has depended since the mid-1990s. In general, specific discourses that attack the Dalai Lama or the concept of independence precede ideological drives that critique non-approved forms of religious behaviour; the former are more likely to persist.

We also identified a minor type of topic that consists of celebratory discourses that emphasise the state's support for the positive role of (reformed or improved) religion in the economy, art or media. These topics are relatively low-frequency but persistent, suggesting that they mark themes in propaganda or rhetoric rather than specific drives.

Overall, the topic model showed that, besides topics that indicate time-specific drives and ongoing core political themes, there are also a large number of topics that we call "maintenance themes" — concepts, arguments, opinions and insistences that are low-frequency, but persist throughout the research period. They provide a kind of sustained intellectual continuo to the peaks and troughs of regulatory and ideological drives. The anti-Dalai and anti-splittist discourses are also of this type, but are far more prominent and virulent in their profile than most such background themes.

To fully understand and interpret a topic, it is necessary to delve into the representative paragraphs that BERTopic assigns to each cluster. Our objective here has been to demonstrate how topic modelling with BERTopic, employing transformer-based numerical encoding of text, can be used effectively to investigate a corpus of Tibetan documents like *Tibet Daily*. By leveraging modern LLMs (such

as the one from Cohere), which support embeddings for Tibetan texts, we can bypass some of the challenges associated with traditional topic modelling methods like LDA. Given the nature of *Tibet Daily* as a source, with its stated role as a government or Party organ, we have interpreted the topics identified by our model as indicators of political campaigns, ideological drives, the dissemination of policy, and the ongoing shaping of public opinion on major issues relating to religion. Nevertheless, our interpretations are limited, particularly because *Tibet Daily* is only one of many official outlets used by the Tibet authorities, and because it reflects provincial-level priorities, not those at a local level. When a political campaign, new policy directive, or organisational imperative for Party and government cadres disappears from sight in the columns of *Tibet Daily*, it may signal not that that campaign or policy has ended, but that it is at that time being implemented throughout the TAR at the local level. In principle, we would therefore aim to expand the sources for our topic model to include local as well as provincial-level media in Tibet, a difficult project. Nevertheless, the use of topic modelling on media at the provincial level provides an abundance of information and insights about the complex, multiform nature of policy implementation and propaganda practices in the TAR.

Bibliography

Barnett, Robert, and Mickey Spiegel.

Cutting Off the Serpent's Head: Tightening Control in Tibet, 1994–1995. London: Tibet Information Network and New York: Human Rights Watch, 1996.

Blei, David M., Andrew Y. Ng, Michael I. Jordan

“Latent Dirichlet Allocation,” *Journal of Machine Learning Research* vol. 3, 2003, pp. 993-1022. Available online at <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf> (accessed January 26, 2025).

Chang Chuan 常川 and Chen Yuejun 陈跃军

“吴英杰：狠抓既定维稳措施落实 确保社会大局和谐稳定” (Wu Yingjie: Pay Close Attention to the Implementation of the Established Stability Maintenance Measures to Ensure the Harmony and Stability of the Overall Social Situation). 西藏日报 (*Tibet Daily*). Posted on 共产党新闻网 [Chinese Communist Party News Network], 26 August, 2020. <http://cpc.people.com.cn/n1/2020/0826/c64102-31837539.html>.

Chang Chuan 常川, Chen Zhiqiang 陈志强 and Chen Yuejun 陈跃军

“西藏自治区代表团向昌都解放纪念碑敬献花篮、看望驻昌 部队官兵、宗教界人士并与各族各界代表座谈 吴英杰讲话” [The Delegation of the Tibet Autonomous Region Presented Flower Baskets to the Chamdo Liberation Monument, Visited the Officers and Soldiers of the Troops Stationed in Changdu, Religious Figures, and Held Discussions with Representatives of All Ethnic Groups and Walks of Life: Wu Yingjie Delivered a Speech]. 西藏日报 (*Tibet Daily*), reposted by cpcnews.cn, 11 October, 2020. Available online at <http://cpc.people.com.cn/n1/2020/1011/c117005-31887459.html> (accessed January 20, 2025).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova.

“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *North American Chapter of the Association for Computational Linguistics*, 2019. Available online at <https://arxiv.org/pdf/1810.04805> (accessed January 15, 2025).

Engels, James, and Robert Barnett

“Developing a Semantic Search Engine for Modern Tibetan”, *Revue d'Etudes Tibétaines* 74, 2025, pp. 262–283.

Goldstein, Melvyn C., Ben Jiao, and Tanzen Lhundrup.

On the Cultural Revolution in Tibet: The Nyemo Incident of 1969. Berkeley, CA: University of California Press, 2009.

Grootendorst, Maarten

“BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure,” *arXiv preprint*, 2022. [doi:10.48550/arXiv.2203.05794](https://doi.org/10.48550/arXiv.2203.05794).

Hartley, Lauran

“Tibetan Publishing in the Early Post-Mao Period,” *Cahiers d’Extrême-Asie* (15), pp. 231-252. [doi:10.3406/asie.2005.1227](https://doi.org/10.3406/asie.2005.1227)

HRW (Human Rights Watch)

“China: New Political Requirements for Tibetan Monastics. Authorities ‘Sinicizing’ Religion,” *Human Rights Watch*, 2018. Available online at <https://www.hrw.org/news/2018/10/30/china-new-political-requirements-tibetan-monastics> (accessed January 15, 2025).

Knierim, Aenne, Michael Achmann, Ulrich Heid and Christian Wolf

“Divergent Discourses: A Comparative Examination of Blackout Tuesday and #BlackLivesMatter on Instagram,” *CLiC-it 2024: Tenth Italian Conference on Computational Linguistics*, 2024. Available online at https://ceur-ws.org/Vol-3878/53_main_long.pdf (accessed January 26, 2025).

Kyogoku, Yuki, Franz Xaver Erhard, James Engels, and Robert Barnett

“LLM in Low-resourced language NLP: Developing a Basic spaCy Modern Tibetan Language Model from Scratch,” *Revue d’Etudes Tibétaines* 74, 2025, pp. 187–220.

MacInnis, Donald E.

Religion in China Today: Policy and Practice. Maryknoll NY: Orbis, 1989.

Meelen, Marieke

“Classical Tibetan Word Embeddings (Version 1),” [Data set]. *Zenodo*, 2022. [doi:10.5281/zenodo.6782247](https://doi.org/10.5281/zenodo.6782247).

Navarretta, Costanza and Hansen, Dorte Haltrup

“According to BERTopic, what do Danish Parties Debate on when they Address Energy and Environment?” In *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*, Association for Computational Linguistics, 2023, pp. 59–

68. Available online at <https://aclanthology.org/2023.cpss-1.6.pdf> (accessed January 26, 2025).

Potter, Pitman B.

“Belief in Control: Regulation of Religion in China,” *The China Quarterly* 174, 2003, pp. 317–337. [doi:10.1017/S0009443903000202](https://doi.org/10.1017/S0009443903000202).

Sabbagh, Christina

“Improving alignment for low-resource parallel corpora”, Master of Science, Speech and Language Processing, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, 2023.

Tsering Wooser.

Forbidden Memory: Tibet during the Cultural Revolution. Lincoln, NA: Potomac Press, 2020.

Xinhuanet

“习近平：全面贯彻新时代党的治藏方略 建设团结富裕文明和谐美丽的社会主义现代化新西藏” [Xi Jinping: Comprehensively Implement the Party’s Strategy for Governing Tibet in the New Era and Build a Socialist Modernised Tibet That is United, Prosperous, Civilised, Harmonious and Beautiful]. *Xinhuanet*, 29 August, 2020. Available online at http://www.xinhuanet.com/politics/leaders/2020-08/29/c_1126428221.htm (accessed January 20, 2025).

Xing, Ying and Wenjing Ni

““Mao Fever” in France: The Reception of Maoism in the French Mass Media, 1963-1979,” *American Journal of Chinese Studies* 31 (1), 2024, pp. 1-24.

Zhang Xiaoming.

China’s Tibet. China Intercontinental Press, 2004.

Zhao Shenying

西藏风云 [Tibet Storm]. Beijing: Xinhua Publishing House, 1987. Available online at <https://books.google.com/books?id=-IPTAAAAMAAJ> (accessed January 25, 2025).

Appendix: A global display of all of the topics

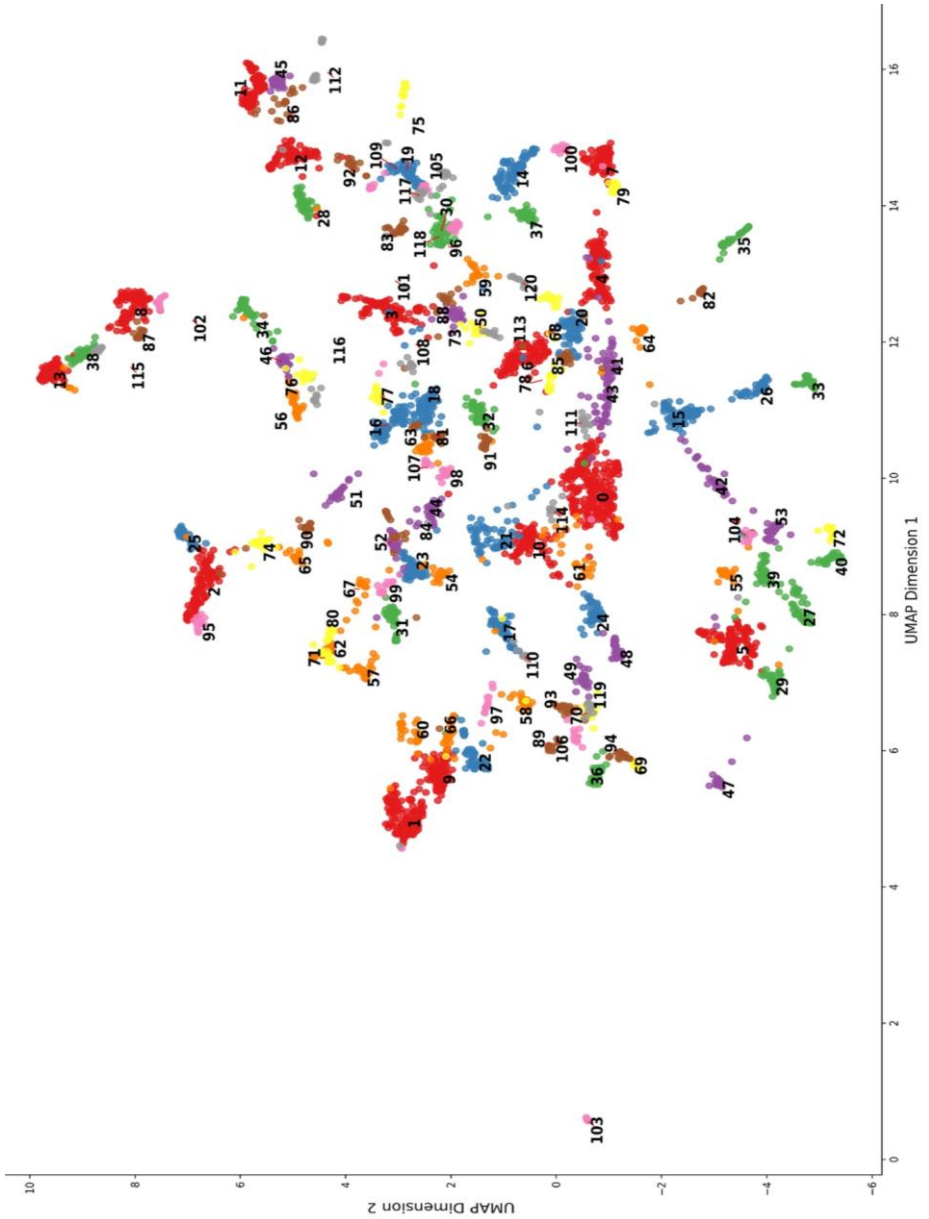


Figure 6 UMAP Projection in 2 dimensions of 121 Topic Clusters Extracted Using BERTopic

The UMAP plot in Figure 6 displays the distribution of all clusters within the latent semantic space. The 121 topics (0-120) are each coloured and numbered. The UMAP algorithm reduces the dimensionality of the 768-dimension vector space to two dimensions for visualisation purposes, ensuring that points that are close in the high-dimensional space remain near each other in the two-dimensional visualisation.¹³



¹³ Topic #103 reports speeches in *Tibet Daily* by Li Kexiang over the entire period. These are not about Tibet, but mention religion (རྗེས་ལུགས་), and thus were included in the subcorpus. In Figure 6 the cluster for Topic #103 is far to the left.