Context is Key(NMF): Modelling Topical Information Dynamics in Chinese Diaspora Media

Ross Deans Kristensen-McLachlan[1,2,*], Rebecca M. M. Hicke[1,3], Márton Kardos[1] and Mette Thunø[4]

[1]*Center for Humanities Computing, Aarhus University, Denmark*
[2]*Department of Linguistics, Cognitive Science, and Semiotics, Aarhus University, Denmark*
[3]*Department of Computer Science, Cornell University, USA*
[4]*Department of Global Studies, Aarhus University, Denmark*

Abstract

Does the People's Republic of China (PRC) interfere with European elections through ethnic Chinese diaspora media? This question forms the basis of an ongoing research project exploring how PRC narratives about European elections are represented in Chinese diaspora media, and thus the objectives of PRC news media manipulation. In order to study diaspora media efficiently and at scale, it is necessary to use techniques derived from quantitative text analysis, such as topic modelling. In this paper, we present a pipeline for studying information dynamics in Chinese media. Firstly, we present KeyNMF, a new approach to static and dynamic topic modelling using transformer-based contextual embedding models. We provide benchmark evaluations to demonstrate that our approach is competitive on a number of Chinese datasets and metrics. Secondly, we integrate KeyNMF with existing methods for describing information dynamics in complex systems. We apply this pipeline to data from five news sites, focusing on the period of time leading up to the 2024 European parliamentary elections. Our methods and results demonstrate the effectiveness of KeyNMF for studying information dynamics in Chinese media and lay groundwork for further work addressing the broader research questions.

Keywords

keywords, novelty, contextual topic models, Chinese, information dynamics

1. Introduction

A number of major elections took place in the West over the course of 2024. Across Europe, citizens took to the polls in early June to elect members to the European Parliament. In France, the election of a new *Assemblée nationale* caused political turmoil, while the United Kingdom voted in a Labour government for the first time in 14 years. On the other side of the Atlantic Ocean, the United States of America will vote to determine their new President in November. The fallout of these elections remains to be seen but it seems clear that this year is one of political change and upheaval.

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

[*] Corresponding author.

rdkm@cc.au.dk (R. D. Kristensen-McLachlan); rmh327@cornell.edu (R. M. M. Hicke); martonkardos@cas.au.dk (M. Kardos); mettethunoe@cas.au.dk (M. Thunø)
https://rmatouschekh.github.io (R. M. M. Hicke)
0000-0001-8714-1911 (R. D. Kristensen-McLachlan); 0009-0006-2074-8376 (R. M. M. Hicke); 0000-0001-9652-4498 (M. Kardos); 0000-0001-9410-2363 (M. Thunø)

Much digital ink is spilled on these topics in Western media as the various electorates determine their preferences before elections and digest the fallout afterwards. Moreover, a significant part of this media coverage is fundamentally persuasive, aiming to convince voters to bet on the candidate who most closely aligns with the social and economic ideology of the media outlets and their owners [13]. Likewise, coverage of these elections is not limited to European media institutions, with media outlets around the world updating their readership on how these elections impact them.

In this context, one particular type of media stands out as especially interesting: ethnic Chinese media targeting diaspora communities in Europe, a group which by some estimates comprises around 1.5-3 million individuals. These media outlets are potentially invaluable sources for understanding how the Chinese government and the Chinese Communist Party (CCP) attempt to influence the diaspora. Furthermore, studying these outlets potentially provides unique insights into how China views itself in relation to the West by showing how the PRC presents itself to its diaspora groups. A growing body of literature has already begun to address these questions in the context of social media [28, 29] or in terms of digital infrastructure more generally [10, 11]. In ongoing research, our aim is to assess whether Chinese diaspora news sources intend to impact opinions on elections in the West during 2024. We attempt to understand the control of information flow in Chinese diaspora media and how this control is used to set specific agendas during electoral periods: promoting certain political parties or individual candidates, polarizing citizens, and attacking or promoting specific political positions.

To pursue this research, we design a pipeline for analyzing large amounts of Chinese-language news data. First, we introduce KeyNMF, a novel approach to creating context-sensitive topics models via transformer-based encoder models. KeyNMF can be trivially applied across different languages and in data scarce environments, and is shown here to create coherent, human-interpretable outputs when working with Chinese language data. We then integrate KeyNMF with existing techniques for describing the information dynamics of complex systems which measure the novelty and resonance of information present in a system over time. We use this pipeline to perform preliminary analysis on our dataset of Chinese diaspora media, finding clear trends in the novelty and resonance signals which correlate with significant political events. The results presented are thus intended to be both a proof of concept and a stepping stone towards more meaningful understanding of the dynamics underlying Chinese diaspora media.

2. Related Work

2.1. Information Dynamics

The study of information dynamics in complex cultural systems has been a central aspect of research in computational humanities and cultural analytics in recent years. One of the most promising approaches to this problem was introduced in [3] which studied the shifting debates which took place during the French Revolution. In this approach, divergence in content between different time slices can be calculated using information-theoretic measures. These measures can then be used to quantify two interrelated values: the *novelty* of the system, or how much the new time slice diverges from preceding time slices; and the *resonance* of this information, which describes how information persists over time.

Novelty-resonance patterns have been studied in a number of different discourse domains. [21] demonstrate their usefulness in identifying so-called trend reservoirs on Reddit. Similar interaction patterns between novelty and resonance have been successfully employed to study the manner in which online news media responded to catastrophic events [18, 20, 19]. In [31], the same fundamental method of analysis demonstrates that novelty-resonance patterns clearly track major social and historical events in the 20th century, using data taken from the front page of Dutch newspapers.

Calculating these underlying dynamics requires the creation of some kind of numerical representation of the data. Specifically, the difference between individual windows is computed by finding the windowed relative entropy, in this case calculated using Jensen-Shannon Divergence (JSD). Since JSD computes the distance between probability distributions, the numerical representations of the data are required to take that form. In [2], this was achieved by calculating the probabilities of a pre-trained, BERT-based emotion classification model, where the predicted probabilities for each label created a distribution over emotions for each document. However, for most purposes, novelty and resonance are calculated based on distributions generated by a probabilistic topic model.

2.2. Vanilla LDA

Typically, novelty and resonance are calculated from topic probability distributions extracted by Latent Dirichlet Allocation (LDA) [9, 7]. Topic distributions in documents are a natural choice for information dynamics, as they are immediately usable with entropy-based measures. LDA is a generative bag-of-words model, which assumes that a document contains a mixture of topics and all words in the document are drawn from this mixture distribution.

However, LDA has a number of well-known shortcomings. Documents have to be heavily pre-processed for optimal results; otherwise, the topic descriptions produced by the model are often contaminated by noise and stop words [16]. In addition, since LDA makes the bag-of-words assumption, it cannot utilize contextual and syntactic information, nor general properties of natural language learned from outside sources. Finally, LDA is sensitive to hyperparameter choices and Wallach, Mimno, and McCallum [30] demonstrate that using symmetric Dirichlet priors, which is the case in canonical implementations [25, 22] and the majority of academic studies, can lead to sub-optimal performance.

There have also been challenges to the generalizability of LDA from the perspective of Chinese NLP, as the primary structural and semantic unit of Chinese is the character rather than the word [32, 24]. While these concerns might be overstated, working with Chinese language data causes specific challenges in terms of tokenization and semantics which directly impact the efficacy of traditional LDA approaches to topic modelling.

2.3. Alternatives to LDA

A major shortcoming of LDA when trying to model change over time is that topics are calculated over all documents, essentially flattening any temporal aspect of the data. This is undesirable, since topics themselves naturally evolve over time, meaning that LDA may not reflect the true dynamics of a system. These issue is partly rectified by dynamic topic models [8] which account for temporal changes in topics with a state-space model. However, Dynamic LDA models are even more parameter-rich than the vanilla implementation and thus amplify its limitations.

Recently, contemporary topic models have shown that it is possible to utilize embeddings from the sentence transformers [27] to infuse contextual information into topic models and to allow for transfer learning [5, 6, 14, 1, 16]. This contextual information can lead to more coherent and semantically interpretable topics. In addition, since these models draw on existing pre-trained language models, they do not require training a generative model from scratch. This means that it is possible to train topic models in data scarce contexts where traditional LDA might perform poorly.

Among the most popular of these contemporary models is BERTopic [14], which also has dynamic modelling capabilities. In this model, topic-term importances are estimated post-hoc on pre-defined time slices based on one underlying topic model. However, as with LDA, BERTopic is sensitive to pre-processing [16]. Additionally, because BERTopic is a clustering topic model, documents are only assigned a single topic label. This renders the model impractical in settings where documents are expected to contain multiple topics and means that BERTopic is not suitable for calculating novelty and resonance, since the entropy calculations assume probability distributions over documents.

3. KeyNMF

We propose KeyNMF, a novel topic modelling approach that utilizes neural text embeddings. KeyNMF builds on the reliability, stability [4], scalability [17], and interpretability of Non-negative Matrix Factorization (NMF) [12], while mitigating its sensitivity to pre-processing and making use of contextual information in texts. This is achieved by: 1) computing keyword importances from documents with contextual embeddings (similar to KeyBERT [15]); and 2) decomposing those importances with NMF.

We release an implementation of KeyNMF as part of the `Turftopic` Python package.[1]

3.1. Model Description

KeyNMF operationalizes topic extraction as the following steps:

1. For each document d:
   a) Let x_d be the document's embedding produced with an encoder model.
   b) Let v_w be the word embedding of a word w produced with the same encoder model.
   c) Let K_d be the set of N keywords in d with the highest cosine similarity to d:

$$K_d = \arg\max_{K^*} \sum_{w \in K^*} \text{sim}(x_d, v_w), \text{ where } |K_d| = N \text{ and } w \in d$$

[1] https://x-tabdeveloping.github.io/turftopic/

2. Arrange the keyword similarities into a non-negative keyword matrix M. Let M_{dw} be the importance of keyword w in document d:

$$M_{dw} = \begin{cases} \text{sim}(d, w), & \text{if } w \in K_d \text{ and } \text{sim}(x_d, v_w) > 0 \\ 0, & \text{otherwise.} \end{cases}$$

3. Decompose M with non-negative matrix factorization: $M \approx WH$, where W is the document-topic matrix, and H is the topic-term-matrix. This is achieved with coordinate-descent, minimizing the square loss $L(W, H) = ||X - WH||^2$.

3.2. Dynamic KeyNMF

KeyNMF can be used for modelling topics' evolution in a corpus over time. This is done by first computing a global model over the entire corpus, then calculating time-specific topic-term importances in predefined time slices. Specifically:

1. Compute the keyword matrix M for the whole corpus.
2. Decompose M with non-negative matrix factorization: $M \approx WH$.
3. For each time slice t:
   a) Let W_t be a subset of W and M_t a subset of M for the documents in time slice t.
   b) Obtain the topic-term-matrix for t with NMF while fixing W_t:

   $$H_t = \underset{H^*}{\arg\min} ||M_t - W_t H^*||^2$$

   c) The temporal importance of topic j is then $I_{tj} = \sum_{d \in t} (W_t)_{dj}$, where all d are documents in time slice t. We can obtain pseudo-topic distributions in the time-slices by L1-normalizing the temporal importances: $\hat{P}_{tj} = \frac{I_{tj}}{\sum_i I_{ti}}$.

Since NMF is not a probabilistic model, we use temporal pseudo-probabilities as a proxy for topic distributions.

3.3. Performance

To demonstrate KeyNMF's effectiveness as a topic model, we evaluate its performance using the `topic-benchmark` Python package and the `paraphrase-multilingual-MiniLM-L12-v2`[2] embedding model. 15 keywords are extracted for each document. Our evaluation procedure is based on that of Kardos, Kostkan, Vermillet, Nielbo, Enevoldsen, and Rocca [16], but, since our intended use case is Chinese news data, we ran the benchmark using the same corpora and pipeline as in our investigations (see Sections 4 and 5). Additionally, we utilized paraphrase-multilingual-MiniLM for measuring external word embedding coherence, instead of an English Word2Vec model. [3]

[2] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

[3] This gives Top2Vec an unfair advantage on this metric as it selects descriptive words based on the same criteria as the metric. C_{ex} scores on Top2Vec should thus be interpreted with caution.

Table 1
KeyNMF's performance on Chinese news data against a number of baselines. Topic descriptions were evaluated on diversity (d), internal (C_{in}) and external (C_{ex}) word embedding coherence.

	chinanews			ihuawen			oushinet			xinozhou			yidali-huarenjie		
Model	d	C_{in}	C_{ex}	d	C_{in}	C_{ex}	d	C_{in}	C_{ex}	d	C_{in}	C_{ex}	d	C_{in}	C_{ex}
KeyNMF	0.93	0.29	0.63	0.91	0.17	0.64	0.84	**0.23**	0.58	0.85	0.26	0.55	0.88	0.52	0.57
S³	0.91	0.16	0.47	0.91	0.11	0.47	0.83	0.12	0.54	0.96	0.17	0.55	0.93	0.46	0.52
Top2Vec	0.78	0.14	**0.71**	0.83	0.10	**0.70**	0.87	0.12	**0.73**	0.86	0.14	**0.71**	0.75	0.46	**0.69**
BERTopic	0.89	**0.31**	0.52	0.89	**0.26**	0.50	0.84	0.23	0.50	0.84	**0.26**	0.52	0.91	**0.57**	0.51
CTM$_{\text{combined}}$	**0.99**	0.27	0.52	**0.99**	0.23	0.51	0.99	0.21	0.51	0.98	0.25	0.51	**0.97**	0.54	0.49
CTM$_{\text{zeroshot}}$	0.99	0.28	0.53	0.99	0.23	0.50	**0.99**	0.22	0.50	**1.00**	0.26	0.51	0.97	0.54	0.51
NMF	0.74	0.27	0.57	0.60	0.18	0.53	0.64	0.18	0.54	0.66	0.18	0.56	0.71	0.49	0.54
LDA	0.61	0.19	0.57	0.53	0.16	0.54	0.41	0.13	0.54	0.48	0.14	0.58	0.57	0.34	0.54

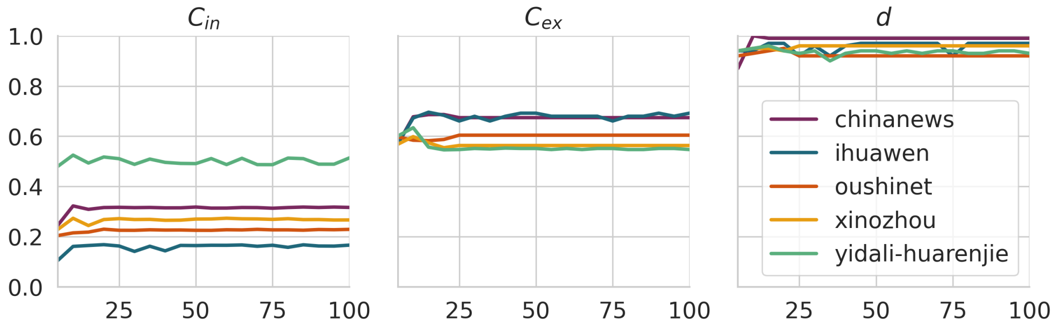

Figure 1: Sensitivity of KeyNMF to the choice of N keywords on multiple metrics and news sources.

Based on our evaluations, KeyNMF's performance is comparable with state-of-the-art contextual topic models, and performs especially well on external coherence, only rivalled by Top2Vec on most corpora, which explicitly selects words based on their proximity in semantic space (see Table 1). The model represents a drastic improvement over classical topic models outperforming both NMF and LDA significantly, indicating that the contextual information infused into the model enhances its performance in a meaningful way.

3.3.1. Sensitivity to Number of Keywords

We additionally test whether the number of keywords extracted from a text influences the model's performance on different corpora, which allows us to determine KeyNMF's robustness to hyperparameter choices. We used the same news sources, pipeline, and quantitative metrics for evaluating this property of the model as for previous evaluations and analyses. The number of keywords was varied from 5 to 100 with a step size of 5 (see Figure 1).

We observed that performance was relatively stable regardless of number of keywords, and converged rather quickly. Only minimal fluctuations are observable with $N > 25$ on most corpora. However, on Xinozhou and Yidali-Huarenjie, lower values of N (5-15) resulted in higher coherence scores. We thus deem 15 keywords a balanced choice of N for further investigations.

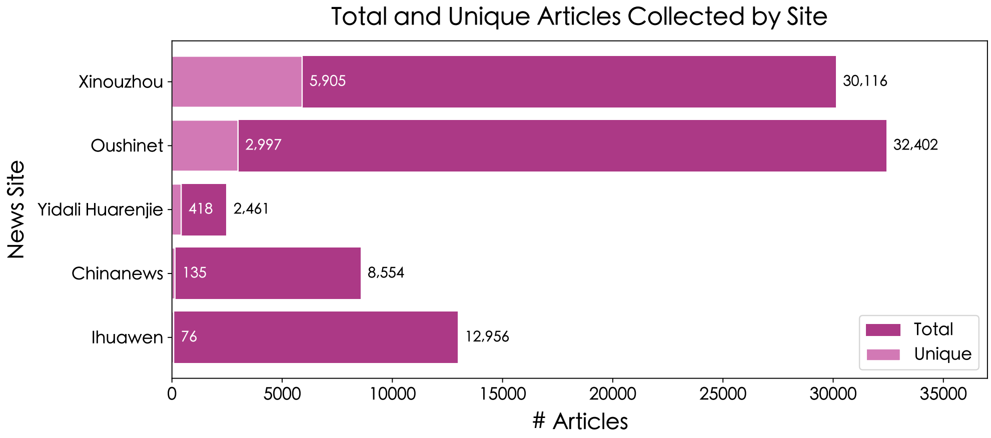

Figure 2: The total and unique number of articles collected for each news site.

4. Data

Having demonstrated the effectiveness of KeyNMF, we use it as the basis for our study of Chinese diaspora media. Our dataset comprises news articles from five sites aimed at Chinese diaspora populations in the EU: Chinanews,[4] Ihuawen,[5] Yidali Huarenjie,[6] Xinouzhou,[7] and Oushinet.[8] We select these sites because they represent a variety of formats, audiences, and perspectives. Oushinet has the largest target audience, with articles in several languages and local journalists writing specifically for the site. In contrast, Xinouzhou reports mostly on Chinese local news, Yidali Huarenjie and Chinanews are community media platforms based in Italy and Scandinavia respectively, and Ihuawen is a weekly magazine based in the United Kingdom.

Our data collection focuses on articles linked from each site's front page and a selection of subpages we deem likely to contain information on international relations, particularly with Europe (listed in Appendix A). We hypothesize that articles linked from these main pages will reflect the topics each news site is attempting to highlight, and will thus provide information on the priorities of the forces backing the media landscape. We scrape all articles linked from each front page and subpage every six hours using a custom web scraper. An article is only scraped once per time point, even if it was linked from multiple pages, but can be scraped multiple times if it appears at multiple time points. Data collection from four sites — Chinanews, Ihuawen, Xinouzhou, and Oushinet — began at 18:15 on April 30, 2024 and collection from the fifth site, Yidali Huarenjie, began at 12:15 on May 7, 2024. Our dataset includes all articles scraped until 6:15 on June 17, 2024, one week after the EU Parliamentary elections took place. Once scraped, we extract the body of each article from the corresponding html file. We attempt to minimize

[4] http://www.chinanews.se
[5] https://ihuawen.com
[6] https://yidali.huarenjie.com
[7] https://www.xinouzhou.com
[8] http://www.oushinet.com

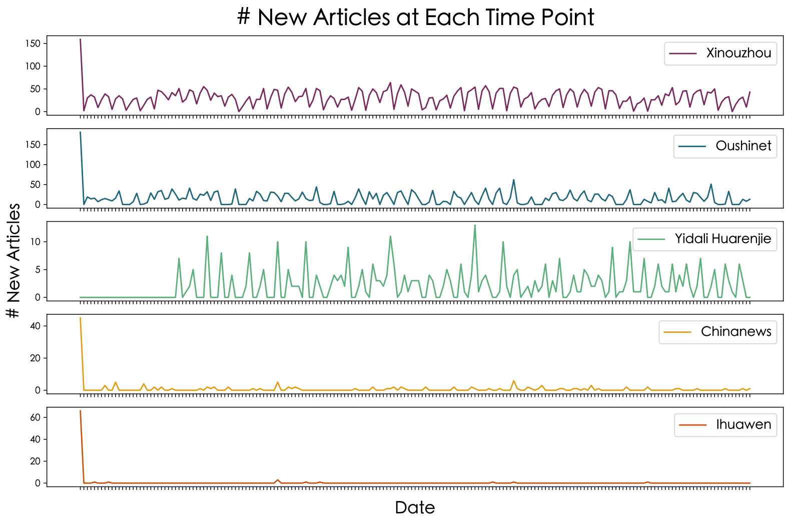

Figure 3: The number of new articles collected at each time point for each source. An article is 'new' if it did not appear in the collected set of articles from the previous time point.

the amount of boilerplate text (e.g. bylines and publication dates) included in the extracted texts; although it is impossible to remove all such text from our dataset, a hand analysis of ten random articles from each news site indicates that the amount of 'junk' text included in the final dataset is minimal.

The total and unique number of articles collected from each site are reported in Figure 2. It is clear that different sites follow different publication patterns. To further validate this, we examine the number of 'new' articles at each time point for each source, or the number of articles that were not included in the last scrape (Figure 3). We see that some sites, like Xinouzhou and Yidali Huarenjie, frequently refresh the articles displayed on their main pages, leading to a larger number of unique articles. In contrast, sites like Ihuawen appear to keep several articles on the main pages for a long time, meaning that they display a very small number of unique articles overall. These differences likely affect the patterns we see in the information systems for each source.

5. Experimental Design

Extracted article texts are embedded with a multilingual transformer-based model [26][9] using the Sentence Transformers library.[10] The embedding is done entirely on a 64-core CPU with 384GB RAM. Each document is embedded once for each time it appears in the dataset. In total, embedding all the documents takes ~2 hours. The maximum sequence length of this embedding model is 128 tokens. Thus, any article longer than 128 tokens is truncated and information from later in the piece is not included in the embedding. Although this is a limitation, we do not consider it prohibitive, as previous research has shown that the bulk of the content in a

[9] `paraphrase-multilingual-MiniLM-L12-v2`
[10] https://sbert.net

news article is presented at the very beginning — a widely-practiced professional standard for journalistic writing known as the *inverted pyramid* [23].

Since our primary interest is understanding the evolution of information dynamics in each news site over time, we use Dynamic KeyNMF to find topic proportions for each timeslice. For keyword extraction, we utilize the `jieba` tokenizer and remove stop-words present in an authoritative list,[11] with the retained tokens then encoded using the same multilingual model as was used on the documents [26]. We fit multiple models with 10, 25, and 50 topics respectively in order to investigate topical dynamics at multiple levels of granularity. Separate models are fit for each news site. The plotted topics over time, top keywords for each topic at each timeslice, and topic distributions at each timeslice are extracted from each model and saved for further analysis.

We then use the topic pseudo-distributions to measure the novelty and resonance signals for each news site and, following [20] and [2], use windowed relative entropy with Jensen-Shannon divergence to calculate both metrics. For a window of size n, the novelty at time point t is the mean entropy of the topic pseudo-distribution at t ($\hat{P}_t$) and the n previous pseudo-distributions. The transience at time point t is the mean entropy of the topic pseudo-distribution at t and the n subsequent pseudo-distributions. Then, the resonance of a time point is the novelty at that point minus the transience. We use a window of size 12 when calculating both signals, which is equivalent to three days of data.

We apply nonlinear adaptive filtering to smooth the extracted novelty and resonance, again following [20] and [2]. This removes noise from the signals by calculating the value at a given time point relative to the surrounding time points. We use a span of 56, the same as [2], for smoothing. The code we use for calculating novelty and resonance is adapted from that released alongside [2] and [20].

6. Results and Discussion

We find clear trends in the novelty and resonance signals that correlate to significant events in the EU during the period studied: Xi Jinping's European Tour (May 5-10), Putin's state visit to China (May 16-17), and the EU parliamentary elections (June 6-9). Our analysis focuses on the novelty and resonance trends extracted from the KeyNMF models with ten topics as these provide the clearest signals. The results for 25 and 50 topics are included in Appendix C.1. We additionally focus our in depth discussion of the results on the two largest news sources, Xinouzhou and Oushinet, for this preliminary validation of the pipeline.

We see spikes in novelty of varying strengths for both Xinouzhou and Oushinet during Xi Jinping's European tour (Figure 4). There are also corresponding dips in resonance before his tour for both sites, followed by increases in resonance during the tour. This indicates that novel information is introduced to the site ecosystems during the tour which replaces previous topics of interest, and which persists in the system for some time.

One of the most productive aspects of Dynamic KeyNMF is that it allows us to study topic fluctuations over time. Thus, we explore which topical shifts contribute to changes in the novelty and resonance signals. For example, on Oushinet, the time period during Xi Jinping's

[11] https://github.com/stopwords-iso/stopwords-zh/blob/master/stopwords-zh.txt

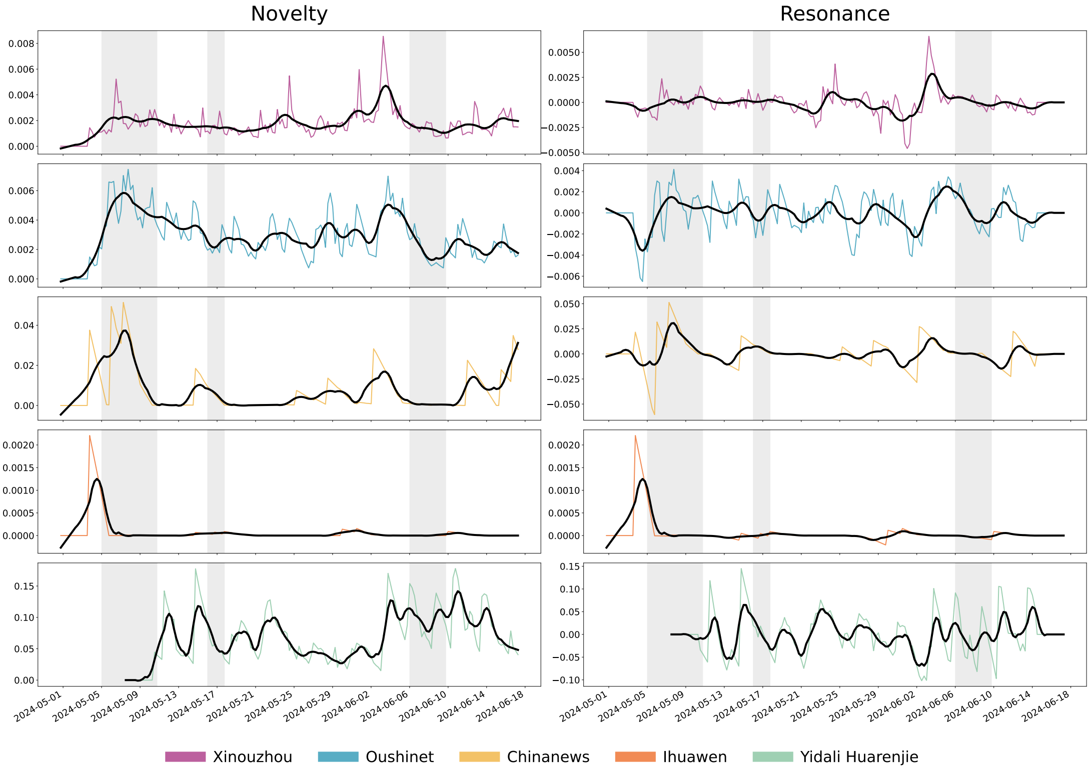

Figure 4: The novelty and resonance plots for each news site from KeyNMF with ten topics. The three shaded areas represent Xi Jinping's European tour (May 5-10, 2024), Putin's state visit to China (May 16-17, 2024), and the EU parliamentary elections (June 6-9, 2024). Note that the y-axis ranges differ for each chart.

European tour is associated with high pseudo-probabilities for a topic defined by the keywords *Paris, France* and *state visit* and a topic defined by *President*, *China*, and *Xi Jinping* (Appendix C.2, Figure 9). Towards the end of the tour, a topic on diplomacy and *bilateral relations* between China and France also gains prominence. For Xinouzhou, this time period contains a peak in the pseudo-probabilities for two topics on Hungary and Chinese relations with Hungary, one of the locations on the tour.

Similarly, there is a noticeable spike in the novelty and resonance for Oushinet directly before Putin's state visit to China. This period is marked by relatively high pseudo-probabilities for a topic characterized by the terms *China*, *Beijing*, *Chinese*, and *Chinese News Service* and a topic with the keywords *Russia*, *Ukraine*, *Putin*, and *Moscow* (Appendix C.2, Figure 7).

Most significantly for this study, there are fluctuations in novelty and resonance for both sites around the EU parliamentary elections. Specifically, there are peaks in the novelty and resonance signals for Xinouzhou and Oushinet before and after the elections, with troughs throughout much of the election period. We hypothesize that these trends reflect a focus on election-related news which begins in early June and continues through the elections and then an introduction of new topics after their end. Again examining the topic distributions, we see that for Oushinet the period before and during the election is marked by high pseudo-probabilities for two topics directly related to the parliamentary elections, one topic surrounding the Spanish prime minister, and two on Russia and Ukraine and the Israel-Palestine war (Appendix C.2, Figure 8). Interestingly, pseudo-probabilities for the topic most directly focused on the elections continued to grow even after the election, suggesting that Oushinet was still discussing the election results during this time. Similarly, for Xinouzhou, three topics focused on the UK elections, Europe broadly, and the Spanish prime minister were comparatively prominent towards the end of May and beginning of June.

Overall, we find that this pipeline allows us to effectively locate changes in news ecosystems, correlate these changes to political and cultural events of interest, and further explore possible reasons for these changes via topic models. It reveals differences in media responses both between events and between sites, while also demonstrating the similarities in sites' news ecosystems, such as the increased discussion of the Spanish prime minister on both Xinouzhou and Oushinet before the EU parliamentary elections. We believe that the combination of the novelty and resonance metrics with the novel KeyNMF topic model will permit further in-depth analysis of these media sites and facilitate research on other Chinese-language domains.

7. Conclusion

In this paper, we present a pipeline designed to facilitate research on the underlying information dynamics of Chinese diaspora media published in Europe. This pipeline combines existing information-theoretic methods that model how new information enters and persists in systems with a novel topic model, KeyNMF. KeyNMF overcomes some of the weaknesses of previous traditional and contextual topic models, demonstrating high performance on standard benchmarks. We validate this pipeline through preliminary experimentation on our dataset of Chinese diaspora media, finding that it reveals informational trends that correlate with major, newsworthy events in European politics and allows for further analysis of the topical changes that cause those trends. While further qualitative research is required to fully understand these dynamics, we believe that we have presented a major step forward in terms of context-sensitive and interpretable topic modelling and information dynamics which can generalize to multilingual and data scarce environments.

Acknowledgments

Part of the computation done for this project was performed on the UCloud interactive HPC system, which is managed by the eScience Center at the University of Southern Denmark.

References

[1] D. Angelov. *Top2Vec: Distributed Representations of Topics.* 2020. arXiv: 2008.09470 [cs.CL].

[2] R. B. Baglini, S. M. Østergaard, S. N. Larsen, and K. L. Nielbo. “Emodynamics: Detecting and Characterizing Pandemic Sentiment Change Points on Danish Twitter”. In: *Proceedings of the Fourth Conference on Computational Humanities Research, CHR 2022*. Antwerp, Belgium, 2022, pp. 162–176.

[3] A. T. J. Barron, J. Huang, R. L. Spang, and S. DeDeo. “Individuals, Institutions, and Innovation in the Debates of the French Revolution”. In: *Proceedings of the National Academy of Sciences* 115.18 (2018), pp. 4607–4612. DOI: 10.1073/pnas.1717729115.

[4] M. Belford, B. Mac Namee, and D. Greene. “Stability of Topic Modeling via Matrix Factorization”. In: *Expert Systems With Applications* 91.1 (2018), pp. 159–169. DOI: 10.1016/j.eswa.2017.08.047.

[5] F. Bianchi, S. Terragni, and D. Hovy. “Pre-Training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online, 2021, pp. 759–766. DOI: 10.18653/v1/2021.acl-short.96.

[6] F. Bianchi, S. Terragni, D. Hovy, D. Nozza, and E. Fersini. “Cross-lingual Contextualized Topic Models with Zero-shot Learning”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online, 2021, pp. 1676–1683. DOI: 10.18653/v1/2021.eacl-main.143.

[7] D. M. Blei. “Probabilistic Topic Models”. In: *Communications of the ACM* 55.4 (2012), pp. 77–84. DOI: 10.1145/2133806.2133826.

[8] D. M. Blei and J. D. Lafferty. “Dynamic Topic Models”. In: *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh, Pennsylvania, USA, 2006, pp. 113–120. DOI: 10.1145/1143844.1143859.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. “Latent Dirichlet Allocation”. In: *Journal of Machine Learning Research* 3.1 (2003), pp. 993–1022. DOI: 10.5555/944919.944937.

[10] V. Brussee. “Authoritarian Design: How the Digital Architecture on China’s Sina Weibo Facilitate Information Control”. In: *Asiascape: Digital Asia* 9.3 (2022), pp. 207–241. DOI: 10.1163/22142312-bja10033.

[11] K. Chan and C. Alden. “<Redirecting> the Diaspora: China’s United Front Work and the Hyperlink Networks of Diasporic Chinese Websites in Cyberspace”. In: *Political Research Exchange* 5.1 (2023), pp. 1–21. DOI: 10.1080/2474736x.2023.2179409.

[12] A. Cichocki and A.-H. Phan. “Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations”. In: *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* E92.a.3 (2009), pp. 708–721. DOI: 10.1587/transfun.E92.A.708.

[13] K. Gatterman, T. M. Meyer, and K. Wurzer. “Who Won the Election? Explaining News Coverage of Election Results in Multi-Party Systems”. In: *European Journal of Political Research* 61.4 (2022), pp. 857–877. DOI: 10.1111/1475-6765.12498.

[14] M. Grootendorst. *BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure.* 2022. DOI: 10.48550/arXiv.2203.05794. arXiv: 2203.05794 [cs.CL].

[15] M. Grootendorst. *KeyBERT: Minimal Keyword Extraction with BERT.* Version v0.3.0. 2020. DOI: 10.5281/zenodo.4461265.

[16] M. Kardos, J. Kostkan, A.-Q. Vermillet, K. Nielbo, K. Enevoldsen, and R. Rocca. *S^3 – Semantic Signal Separation.* 2024. DOI: 10.48550/arXiv.2406.09556. arXiv: 2406.09556 [cs.LG].

[17] A. Lefèvre, F. Bach, and C. Févotte. "Online Algorithms for Nonnegative Matrix Factorization with the Itakura-Saito Divergence". In: *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).* New Paltz, NY, USA, 2011, pp. 313–316. DOI: 10.1109/aspaa.2011.6082314.

[18] K. L. Nielbo, R. B. Baglini, P. B. Vahlstrup, K. C. Enevoldsen, A. Bechmann, and A. Roepstorff. "News Information Decoupling: An Information Signature of Catastrophes in Legacy News Media". In: *Proceedings of the 2020 European Association for Digital Humanities Conference.* Krasnoyarsk, Russia, 2021, pp. 1–8. DOI: 10.48550/arXiv.2101.02956.

[19] K. L. Nielbo, K. Enevoldsen, R. Baglini, E. Fano, A. Roepstorff, and J. Gao. "Pandemic News Information Uncertainty –News Dynamics Mirror Differential Response Strategies to COVID-19". In: *Plos One* 18.1 (2023), e0278098. DOI: 10.1371/journal.pone.0278098.

[20] K. L. Nielbo, F. Haestrup, K. C. Enevoldsen, P. B. Vahlstrup, R. B. Baglini, and A. Roepstorff. *When No News is Bad News – Detection of Negative Events from News Media Content.* 2021. DOI: 10.48550/arXiv.2102.06505. arXiv: 2102.06505 [cs.CY].

[21] K. L. Nielbo, P. B. Vahlstrup, A. Bechmann, and J. Gao. "Trend Reservoir Detection: Minimal Persistence and Resonant Behavior of Trends in Social Media". In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020).* Amsterdam, the Netherlands, 2020, pp. 290–297. DOI: 10.48550/arXiv.2109.08589.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.1 (2011), pp. 2825–2830.

[23] H. Pöttker. "News and Its Communicative Quality: the Inverted Pyramid – When and Why Did It Appear?" In: *Journalism Studies* 4.4 (2003), pp. 501–511. DOI: 10.1080/1461670032000136596.

[24] Z. Qin, Y. Cong, and T. Wan. "Topic modeling of Chinese language beyond a bag-of-words". In: *Computer Speech & Language* 40 (2016), pp. 60–78. DOI: https://doi.org/10.1016/j.csl.2016.03.004.

[25] R. Řehůřek and P. Sojka. "Software Framework for Topic Modelling with Large Corpora". In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.* Valletta, Malta, 2010, pp. 45–50.

[26] N. Reimers and I. Gurevych. “Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Online, 2020, pp. 4512–4525. DOI: 10.48550/arXiv.2004.09813.

[27] N. Reimers and I. Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Hong Kong, China, 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410.

[28] M. Schliebs, H. Bailey, J. Bright, and P. N. Howard. *China’s Public Diplomacy Operations: Understanding Engagement and Inauthentic Amplification of PRC Diplomats on Facebook and Twitter.* Tech. rep. Oxford, UK: Programme on Democracy & Technology, 2021.

[29] M. Thunø and K. L. Nielbo. “The Initial Digitalization of Chinese Diplomacy (2019–2021): Establishing Global Communication Networks on Twitter”. In: *Journal of Contemporary China* 33.146 (2024), pp. 244–266. DOI: 10.1080/10670564.2023.2195811.

[30] H. M. Wallach, D. Mimno, and A. McCallum. “Rethinking LDA: Why Priors Matter”. In: *Advances in Neural Information Processing Systems*. Vancouver, Canada, 2009, pp. 1–9.

[31] M. Wevers, J. Kostkan, and K. L. Nielbo. “Event Flow – How Events Shaped the Flow of the News, 1950-1995”. In: *Proceedings of the Third Conference on Computational Humanities Research, CHR 2021*. Amsterdam, the Netherlands, 2021, pp. 62–76. DOI: 10.48550/arXiv.2109.08589.

[32] Q. Zhao, Z. Qin, and T. Wan. “Topic Modeling of Chinese Language Using Character-Word Relations”. In: *Neural Information Processing*. Berlin, Heidelberg, 2011, pp. 139–147.

A. News Site Subpages

The subpages scraped for each news site are listed below:

- **Xinouzhou:** France, Italy, Spain, UK, Germany, Hungary, International
- **Ihuawen:** News, Comments & Opinions
- **Oushinet:** Europe, Europe: Germany, Europe: Central and Eastern Europe, Europe: Italy, Europe: Spain, Europe: Other, France, Europe and China, Overseas Chinese community, China, International, Opinion on public affairs
- **Chinanews:** Nordic headlines, China news, Mutual learning among civilizations, Overseas Chinese community, Nordic Commercial Bridge, Overseas thoughts
- **Yidali Huarenjie:** ∅

B. NPMI Coherence

Since NPMI Coherence has historical significance in topic modeling literature, we also evaluated topic descriptions with this metrics. Due to theoretical and practical limitations[16], however, we do not consider NPMI Coherence a good metric for evaluating topic models. For the sake of completeness, we report C_{NPMI} scores in Table 2.

Table 2
C_{NPMI} coherence of different topic models on the studied corpora.

Model	chinanews	ihuawen	oushinet	xinozhou	yidali-huarenjie
BERTopic	0.06	**0.11**	0.08	**0.10**	0.08
CombinedTM	-0.07	-0.02	-0.07	-0.02	-0.12
KeyNMF	-0.21	-0.23	-0.00	-0.04	-0.14
LDA	-0.02	-0.02	0.02	0.03	0.00
NMF	**0.11**	0.08	**0.10**	**0.10**	**0.10**
S^3	-0.37	-0.37	-0.22	-0.18	-0.37
Top2Vec	-0.36	-0.36	-0.25	-0.22	-0.36
ZeroShotTM	-0.07	-0.03	-0.06	-0.03	-0.11

C. Additional Experimental Results

C.1. Novelty and Resonance Ablations

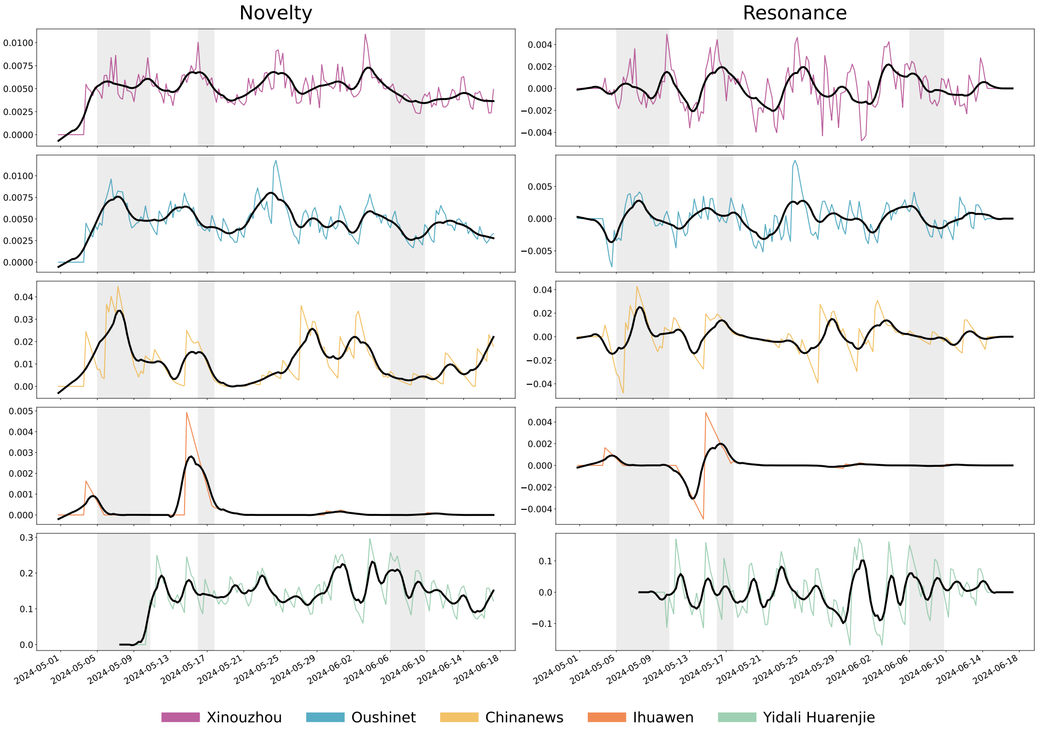

Figure 5: The novelty and resonance plots for each news site from KeyNMF with 25 topics. The three shaded areas represent Xi Jinping's European tour (May 5-10, 2024), Putin's state visit to China (May 16-17, 2024), and the EU parliamentary elections (June 6-9, 2024). Note that the y-axis ranges differ for each chart.

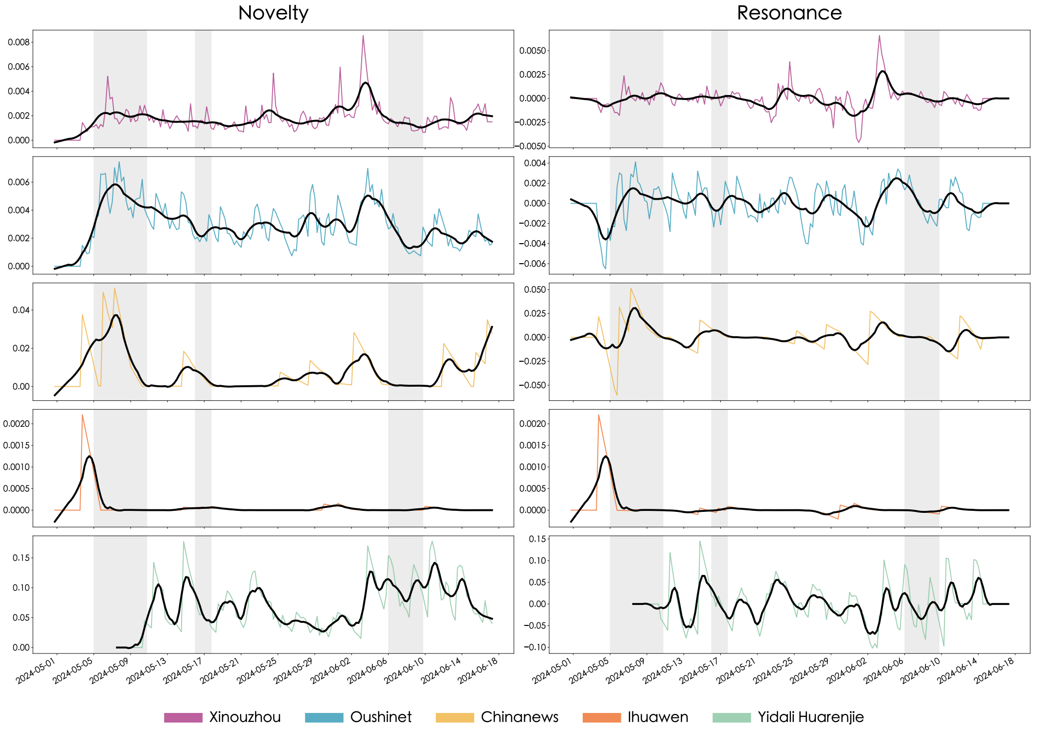

Figure 6: The novelty and resonance plots for each news site from KeyNMF with 50 topics. The three shaded areas represent Xi Jinping's European tour (May 5-10, 2024), Putin's state visit to China (May 16-17, 2024), and the EU parliamentary elections (June 6-9, 2024). Note that the y-axis ranges differ for each chart.

C.2. Topic Distributions Over Time

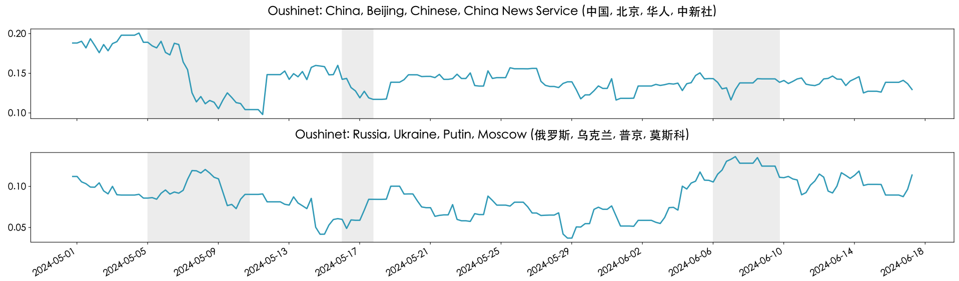

Figure 7: The distributions over time for two topics with high pseudo-probabilities before Putin's state visit to China. These topics are generated by the 10-topic KeyNMF model for Oushinet. Note that the y-axis scale differs for each subplot.

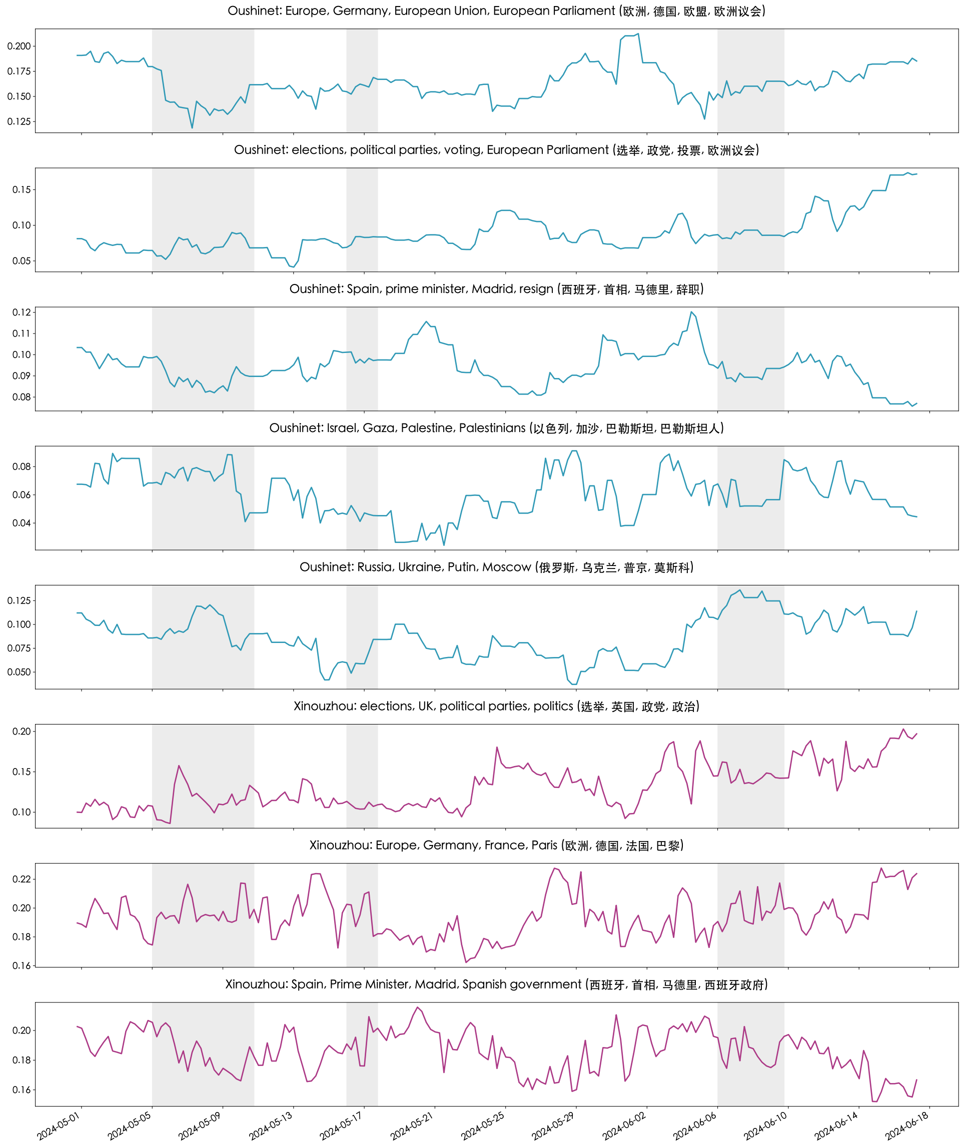

Figure 8: The distributions over time for eight topics with high pseudo-probabilities around the EU parliamentary elections. These topics are generated by the 10-topic KeyNMF models for Oushinet and Xinouzhou. Note that the y-axis scale differs for each subplot.

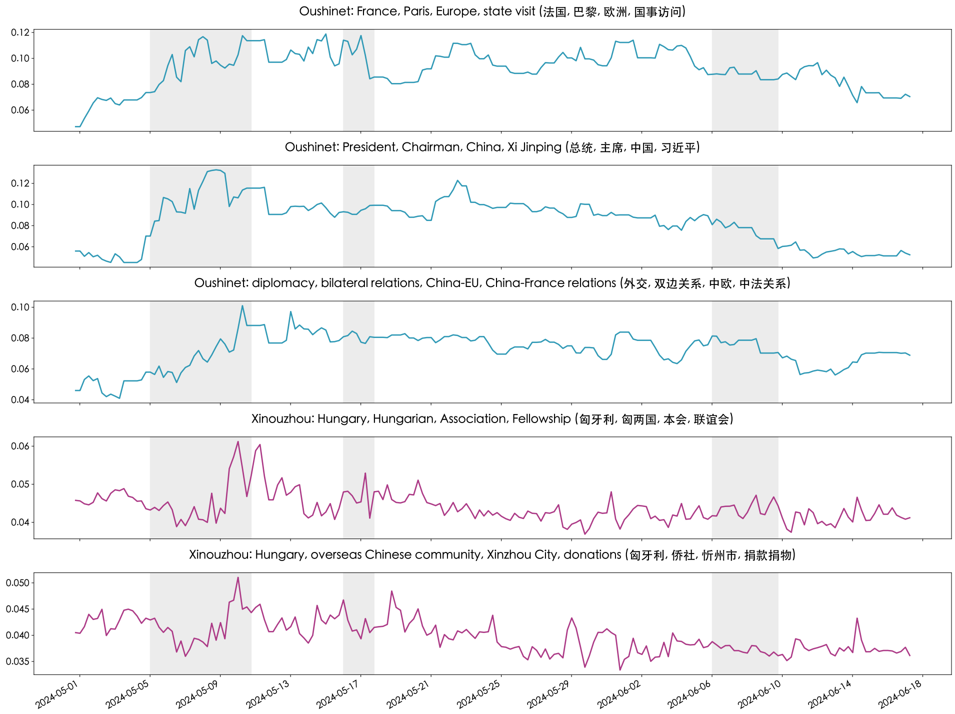

Figure 9: The distributions over time for five topics with high pseudo-probabilities during Xi Jinping's European tour. These topics are generated by the 10-topic KeyNMF models for Oushinet and Xinouzhou. Note that the y-axis scale differs for each subplot.