A Co-concerned Multilingual Topic Detection Model Based on mT5 and Frequency Entropy

Anonymous ACL submission

Abstract

Topic models play a crucial role in various fields such as text classification and semantic extraction. However, the enhancement of the quality of topic words faces persistent challenges and has been explored for a long time. Among these, attention bias stemming from different language cultures often emerges, particularly in hot events. While topic models excel at detecting incident topics, they are susceptible to the influence of bias misguidance. Furthermore, existing topic models encounter limitations when applied to multilingual corpora, as synonymous multilingual representations may disproportionately occupy the forefront of the output sequence. In light of these issues, we propose a model that combines the text clustering algorithm of BERTopic with the extraction of topic words using a tuned mT5. The output words are filtered using a word table that stores words with high information entropy in multiple languages. We conducted experiments on our dataset, demonstrating high performance not only in commonly focused multilingual topic detection but also in the elimination of output redundancy.

1 Introduction

Various language cultures exhibit inherent biases in their focus on daily hot news(Smith et al., 2018). As illustrated in Figure 1, Arabs may demonstrate a heightened concern for religious topics, whereas Americans might prioritize human rights. Distinctive topic preferences are ubiquitous worldwide during specific periods. Detecting common concerns can mitigate the impact of cognitive biases originating from different cultures, offering substantial potential value.

However, existing topic models cannot effectively highlight commonly concerned topics. For instance, LDA(Blei et al., 2003) is a probabilistic generation model capable of detecting topic structures from text using the Dirichlet distribution. HDP(Teh et al., 2004) introduces an infinite probabilistic process to automatically infer topic structures. BERTopic(Grootendorst, 2022) employs an architecture that creates sentence embeddings through SBERT(Reimers and Gurevych, 2019) and utilizes c-tf-idf to extract topic words from clustered texts. None of these models can adequately emphasize co-concerned topics across multiple languages. Furthermore, many keywords with similar meanings often co-occur simultaneously in outputs(Grootendorst, 2022), given their comparable frequency or distribution in the corpus.

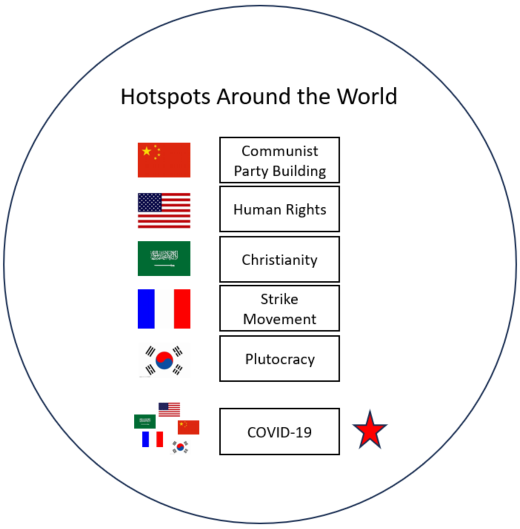


Figure 1: People's attention varies across different countries, leading to attention biases in different languages. The co-concerned topic holds significant value.

With the word sorting function in each topic model, these keywords might occupy a disproportionately large portion of the sequence's front space, causing potentially valuable words to be ranked in a more arbitrary manner behind them. These factors not only perpetuate the influence of attentional bias on topic models but also contribute to low-quality output.

To filter biased topics and extract qualified topic words, we incorporate information entropy to identify co-concerned multilingual topics. The entropy of each word's appearance distribution in different languages is computed as frequency entropy, and synonyms are merged by calculating the cosine similarity of the mBERT-generated vectors. This process aims to create our proposed co-concerned word table.

Subsequently, we utilize the fine-tuned mT5 (Xue et al., 2020) to extract topic words from individual "text" to avoid output loss caused by sorting words from entire "texts." Synonyms can be stored using any desired language form for translation.

To assess the limitations on multilingual corpora, we also evaluate classical topic models using the OCTIS(Terragni et al., 2021) framework. Our method effectively enhances the quality of topic words and the capacity for detecting co-concerned topics. The primary contributions of this paper are:

1. We introduced a new multilingual topic detection model capable of constructing a co-concerning table for effective co-concerned topic detection and removing synonym redundancy.

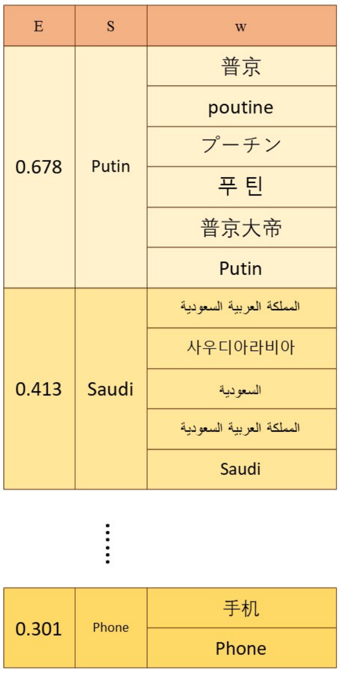
2. We tuned mT5 for extracting from a single text to preserve potential topic words and unveil more details of incidents.

3. The proposed model has been evaluated on our customized datasets, and it outperforms other matrix-factorization based or Neural Network based models in co-concerned degree, thereby validating the effectiveness of leveraging words' frequency entropy under different languages.

2 Related Work

Multilingual topic detection is extensively explored in the domains of information mining and intelligence analysis. In a comprehensive evaluation conducted by Huber et al.(Huber and Spiliopoulou, 2019), there are three methods including neural networks, attention mechanism based and rule-based hybrid method assessed for creating clusters from either multilingual word embeddings or monolingual clusters. Additionally, an ontology-based methodology is proposed for automatic topic detection without prior information, leveraging hierarchical clustering algorithms and a multilingual knowledge base(Gutiérrez-Batista et al., 2018).

Another approach involves detecting underlying topics in multilingual datasets through clustering. This method relies on multilingual aligned embeddings and community detection in (Stefanovitch et al., 2023). Furthermore, the advent of multilingual pretraining models has introduced novel approaches to multilingual topic analysis. The clustering process of BERTopic, upon which our work is based, utilizes multilingual sentence BERT for embedding. Although redundancy can be addressed theoretically by applying maximal marginal relevance (Carbonell and Goldstein, 1998) to the top n words, it hasn't been explored detailedly. Hence, 'diversity' in MMR serves as a hyperparameter experimented within our work.



E	S	w
0.678	Putin	普京
		poutine
		プーチン
		푸 틴
		普京大帝
		Putin
0.413	Saudi	المملكة العربية السعودية
		사우디아라비아
		السعودية
		المملكة العربية السعودية
		Saudi
⋮		
0.301	Phone	手机
		Phone

Figure 2: Here is an illustration of a co-concerned topic word sorting table. Let w represent the words appearing in the corpus, where W denotes the related synonyms in English contained within S. Each item in W is ranked based on the information entropy E_W.

Moreover, given that multilingual sentence embeddings effectively preserve word semantics, the tensor similarity of BERT embeddings is multiplied by each topic's LDA probability value to explore the evolution of topics in (Xie et al., 2020).

While information entropy is commonly employed for keyword extraction, its application in topic clustering is less frequent. Yang et al.(Yang et al., 2013) propose a novel metric to evaluate and rank the relevance of words in a text. The method utilizes the Shannon's entropy difference between the intrinsic and extrinsic modes, empha-

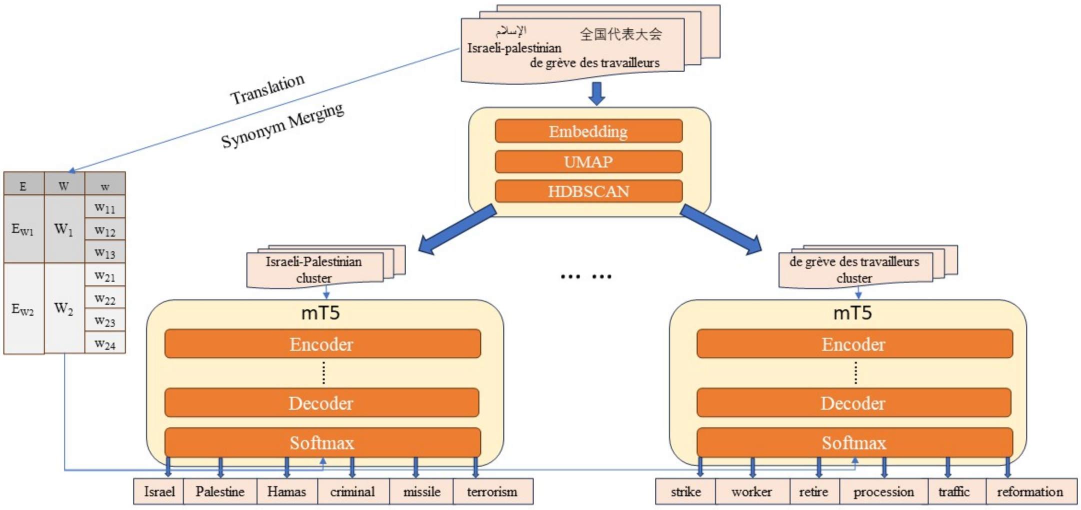


Figure 3: The architecture overview of our model. We build the co-concerned word table sorting words by E and make softmax in mT5 choose output words from it. Topic clustering parts follow BERTopic.

sizing that relevant words significantly reflect the author's writing intention. Singhal et al.(Singhal and Sharma, 2021) introduce and analyze a domain-independent statistical method for keyword extraction using Rényi entropy. Experimental results indicate that the Rényi entropy-based word ranking metric exhibits reliable performance and coherence with previously defined entropy-based methods.

In another study, Xu et al.(Xu et al., 2022) employ link prediction and structural-entropy methods to predict scientific breakthrough topics. The temporal changes in the structural entropy of a knowledge network are utilized to identify potential breakthrough topics. All these works leverage information entropy to unveil potential statistical characteristics of the data, which enlighten us on multilingual commonly concerned topic extraction..

3 Model

3.1 Frequency Entropy Under Different Languages

Shannon's entropy of word frequency under different languages is utilized to construct the Co-concerned words table and rank its entries. Let's assume that the i-th topic contains words as $Topic_n = \{w_{n1}, w_{n2}, \cdots w_{nj}\}$. Following the table-building process of translation and synonym merging, each word w_{ni} is mapped to a specific language form of its synonym W_{ni}, present in the set $S_n = \{W_{n1}, W_{n2}, \cdots W_{nq}\}$.For the merged W_{ni}, we count the frequency of each synonym w_n under language l as F_{lW}. Consequently, the probability of the appearance of W's synonym under language l is:

$$P_{lW} = \frac{F_{lW}}{\sum_{i=1}^{L} F_{iW}} \tag{1}$$

L is the number of languages. Then the frequency entropy E_W for W under each language is given by:

$$E_W = -\sum_{l=1}^{L} P_{lW} \log P_{lW} \tag{2}$$

By ranking the information entropy E_W, we can construct the co-concerned topic word sorting table $T_{co-concerned}$, as illustrated in Figure 2.

3.2 Tuning of mT5

mT5 (Xue et al., 2020) is an encoder-decoder language model, parametrized as $p_\phi(y|x)$, where x represents the single text, and y denotes the generated topic words. Let z be the concatenation of x and y, and h_i represents the activation at time step i. X_{idx} and Y_{idx} denote the sequences of indices corresponding to x and y. The initialization of ϕ utilizes the pretrained mT5_multilingual_XLSum model.To address parts shorter than the maximum input and output length, we add zero to the mask and -100 to the label. This ensures that attention and cross-entropy loss disregard these parts. Subsequently, we perform gradient updates on the following objective:

"Text-Words" Dataset	Zh	Ar	En	Ja	Ko	Mixed	Total
Training Set	990	1000	1000	1034	983	1000	6007
Test Set	379	406	400	403	400	400	2388

Table 1: Distribution of dataset sizes under five languages for mT5's tuning, comprising pure texts and randomly language-mixed texts for each of the five languages.

$$\max_{\phi} \; log\, p_{\phi}(y|x) = \sum_{i \in y_{idx}} log\, p_{\phi}(z_i \mid h_{<i}) \quad (3)$$

As one of the most widely used parameter-efficient tuning methods, prefix-tuning(Li and Liang, 2021) employs continuous word embeddings to optimize prompts instead of discrete tokens. The trainable matrix P_θ is initialized to store the prefix parameters. To prevent unstable optimization and mitigate a slight drop in performance, the matrix undergoes reparametrization, given by $P_\theta[i,:] = \text{MLP}_\theta\left(P'_\theta[i,:]\right)$, where P'_θ is composed of a large feedforward neural network (MLP). Only the prefix parameters θ are trained.

3.3 The Co-concerned Topic Detection Model

UMAP(McInnes et al., 2018) and HDBSCAN(McInnes et al., 2017) are implemented to cluster documents in BERTopic. Each document are fed into mT5 then. Except frequency entropy, we use Google translation and synonym aggregation based on mBERT and cosine similarity to build $T_{co-concerned}$ as where softmax get the output words from.

UMAP(McInnes et al., 2018) and HDBSCAN(McInnes et al., 2017) are implemented to cluster documents in BERTopic. Each document is subsequently processed through mT5. In addition to frequency entropy, the construction of $T_{co-concerned}$ involves Google translation and synonym aggregation based on mBERT, utilizing cosine similarity. The softmax function is applied to obtain the output words from $T_{co-concerned}$.

4 Experiment

Our experiments consist of three parts: mT5 tuning, baseline evaluation, and model evaluation.

4.1 Dataset

Tuning Dataset. The dataset utilized for mT5 tuning is in "text-words" format. Initially, text is collected from news websites in five different languages. Then we sample 30 texts in each batch for BERTopic, applying tf-idf and maximal marginal relevance to obtain high-quality topic words. Additionally, we get one thousand texts that are a random mixture of the five languages, translated into English for extracting topic words (to avoid the inadaptability of tf-idf on multiple languages). Then we mix their corresponding original representations to construct language-mixed "Text-Words" for the dataset. This dataset is further divided into training and test sets, with its configuration detailed in Table 1.

Evaluation Dataset. A customized dataset is essential for evaluating the newly proposed "co-concerning" matrix. For each news topic (Sports, Warfare, Sci. & Tech, Economy, Political), we associate each language with a specific event and introduce a mixed-language event as well, labeled as co-concerned (as illustrated in Table 4). There are five news topics, and each topic comprises 20 event-specific news articles for each language, along with the co-concerned part.

4.2 Tuning on Topic Extraction Task

Topic word extraction task is based on the mT5_multilingual_XLSum model from huggingface with fine-tuning and prefix-tuning methods. We restrict the input length to 1024 tokens and output length to 64 tokens, which is about ten words' output for each text. Both methods implement the optimizer of AdamW with a learning rate of $2e^{-5}$. The learning rate optimization strategy adopts linear strategy. The loss function is cross-entropy, and there are 20 epochs of training. Since we focus on the existence of each single word in the label sequence, the evaluation metric of topic extraction task is rouge-1.

4.3 Tuning on Topic Extraction Task

The topic word extraction task is performed using the mT5_multilingual_XLSum model from Hugging Face, employing fine-tuning and prefix-tuning methods. We constrain the input length to 1024 tokens and the output length to 64 tokens, which equates to approximately ten words' output for each text. Both methods utilize the AdamW opti-

Topic Model	Topic Coherence	Topic Diversity
BERTopic (Best Score)	0.515	0.925
CTM (Best Scorhe)	0.486	0.953
HDP	0.103	0.825
LDA	0.366	0.903
NeuralLDA	0.389	0.893
NMF	0.345	0.509
ProdLDA	0.334	0.855
Ours	0.557	0.961

Table 2: Evaluation results for different models which indicate that our proposed model achieves the highest topic coherence and topic diversity score.

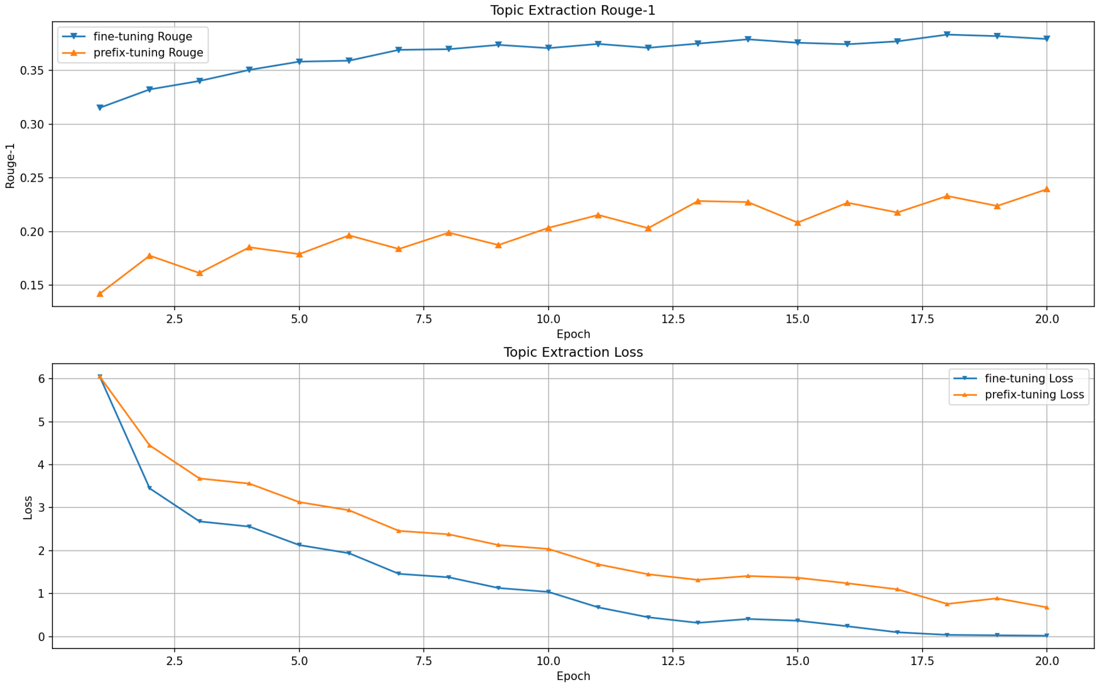


Figure 4: Results of tuning on mT5. We choose rouge-1 as evaluation matrix because only the existence of single word from output is considered. Results of tuning on mT5. We choose rouge-1 as the evaluation metric since only the existence of single word from output is needed to be considered.

mizer with a learning rate of $2e^{-5}$. The learning rate optimization strategy follows a linear approach. The loss function employed is cross-entropy, and the training spans 20 epochs. Given our emphasis on the presence of each individual word in the label sequence, the evaluation metric for the topic extraction task is rouge-1.

4.4 Evaluations

Baselines. We conduct training for BERTopic, CTM(Song et al., 2020), LDA, HDP, NeuralLDA(Srivastava and Sutton, 2017), NMF(Wang and Zhang, 2012), and ProdaLDA(Srivastava and Sutton, 2017) on the OCTIS framework(Terragni et al., UMAP2021). The model hyperparameters are summarized in Table 3(Egger and Yu, 2022). Texts undergo cleaning, and news articles with fewer than 100 words are filtered out. Following the approach in (Abdelrazek et al., 2022), we utilize multiple pretrained multilingual SBERT models for BERTopic and CTM. Parameters not mentioned in Table 3 are set to default values.

The performance of the topic models in this study is assessed using two widely-adopted metrics: topic coherence and topic diversity. For each topic model, the topic coherence is evaluated using the coefficient of variance (C_V(Abdi, 2010)). This metric ranges from [-1, 1], with 1 indicating a perfect association. Topic diversity, as defined by (Dieng et al., 2020), represents the percentage of unique words across all topics. This measure ranges from [0, 1], where 0 suggests redundant topics and 1 signifies more diverse topics.

Recognizing that the evaluation metrics are not suitable for multiple languages, we translate both the output words and input text into English for evaluation. Each model undergoes 200 training runs, and the median value is selected as the final score. The model demonstrating the best performance is considered the baseline.

Co-concerning Measure. The matrix $P^{n}_{co\_concerned}$ gauges the degree of alignment between topic words and the co-concerned event. We consider the top n% of words from each output iteration, denoted by set S. Here L represents the number of languages. For every word W_i, assuming it appears in l_i different languages, and each synonym w is associated with it, we have:

$$P^{n}_{co\_concerned} = \frac{\sum_{i=1}^{n\%*|S|} l_i}{L * n\% * |S|} \tag{4}$$

Given that stop words may get high values, it is imperative to exclude them during the preprocessing stage. Our experiment focuses on the top n% of 200 output words.

Redundancy. Redundancy elimination is a key innovation of our model. The comparative experiment is conducted as follows: assume $S_i^{'}$ represent the i-th set of redundant words from the output, with a total count of n. We tally the number of distinct words within the output set S and then calculate the proportion relative to the length of the original output S using the following formula:

$$R_S = 1 - \frac{|S - S_1^{'} \cup S_2^{'} \cdots \cup S_n^{'}| + n}{|S|} \tag{5}$$

We use 1 as the minuend to obtain the redundancy degree of the output word sequence S.. In each evaluation, we consider 200 words and record the redundancy in the top n% words. The size of the redundant word set depends on the cosine similarity threshold of the word vectors.

5 Results

In this section, we present the results of mT5's tuning and the performance comparison with baseline.

5.1 Tuning Tasks

The tuning results are depicted in Figure 4. Prefix-tuning, with fewer parameters, exhibits significantly inferior performance compared to fine-tuning.. After 20 epochs, the final rouge-1 score is 0.383. Considering synonyms in the result would yield a higher score, implying that mT5 already possesses topic extraction capabilities.

5.2 Model Performance

TC & TD. As shown in Table 5 and Table 6, CTM and BERTopic tried different multilingual embedding model. The best scores is compared with other models as shown in Table 2. Our model has competitive performance with 0.577 coherence and 0.961 diversity. It indicates that as the topic word extraction is from each single text, mT5 can dig enough high performance(TC and TD) words. And the front-ranking ones by frequency entropy tend to be these high quality words.

As illustrated in Table 5 and Table 6, CTM and BERTopic experimented with various multilingual embedding models. The best scores are compared with other models, as presented in Table 2. Our model exhibits competitive performance with a coherence score of 0.577 and a diversity score of 0.961. This suggests that by extracting topic words from individual texts, mT5 can yield sufficiently high-performing words in terms of Topic Coherence (TC) and Topic Diversity (TD).

Co-concerning. As shown in Figure 5, frequency entropy makes about 56% front of words are all five languages co-concerned focus ones. The lower proportion in the back may be due to the fact that the rear words of the table appear in only four or fewer languages. BERTopic's output is ranked by c-TF-IDF, which makes the co-concerned words just appear randomly with no highlighting.

As shown in Figure 5, frequency entropy makes approximately 56% of the words at the forefront are topics of co-concern across all five languages. The lower proportion towards the back may be attributed to words that appear in only four or fewer languages. In contrast, BERTopic's output ranked by c-TF-IDF, presents co-concerned words randomly with no highlighting.

Redundancy. In Figure 6, BERTopic shows much higher proportion of synonyms redundancy in the sequence's front. Our model has 40% less redundancy in the top 10% words and as the statistical vocabulary increases, the proportion of both

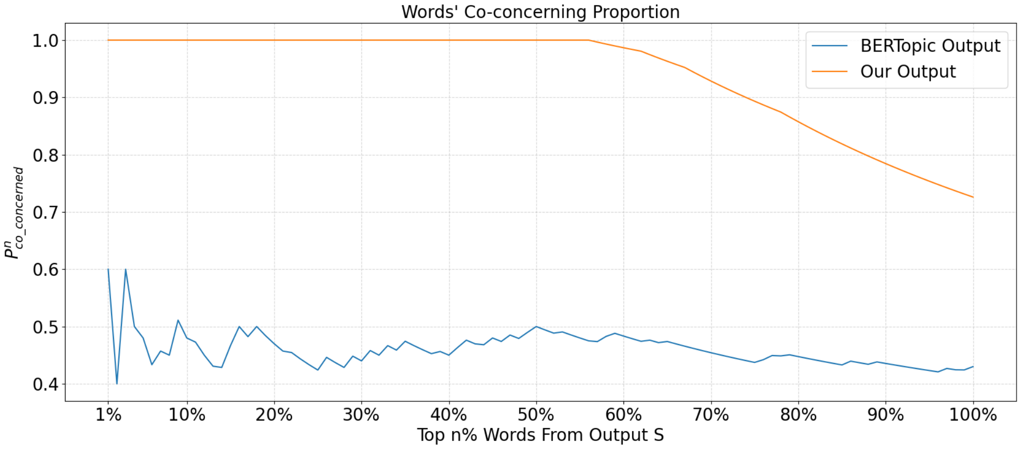


Figure 5: The proportion of co-concerning words in the top n% calculated by Formula 4

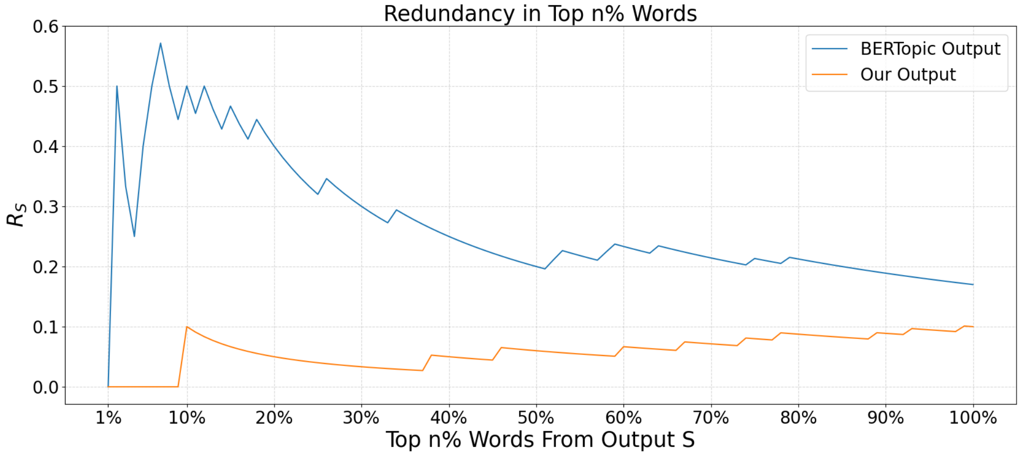


Figure 6: The proportion of synonyms in the top n% words calculated by Formula 5

tends to flatten out. After adding the word sorting function by Table $T_{co-concerned}$, redundant words just appear several times in our model's output, which is the normal efficiency limitation of word vectors generated by mBERT.

"In Figure 6, BERTopic exhibits a notably higher proportion of synonym redundancy in the sequence's front. Our model, on the other hand, demonstrates a 40% reduction in redundancy within the top 10% of words. As the statistical vocabulary expands, the proportions for both models tend to level out. The incorporation of the word sorting function by Table $T_{co-concerned}$ in our model results in redundant words appearing only a few times in the output, which is the inherent efficiency limitation of word vectors generated by mBERT.

6 Conclusion

We proposed a model capable of extracting co-concerned topic words from different languages. Following the BERTopic approach, we clustered texts and utilized mT5 to extract topics from each text. We constructed a word table (referred to as $T_{co-concerned}$ in this paper) to rank the top n words based on the entropy of word appearance frequency in different languages. The merging of synonyms during output generation helps to eliminate redundancy. Experimental results and outputs obtained on our test set (as shown in Table 7) demonstrate that our model significantly outperforms the baseline model, BERTopic. This validates that the performance of multilingual co-concerned topic detection can be effectively enhanced by considering multilingual frequency entropy.

Regarding the limitations, our model necessitates refinement in constructing the word table due to the influence of the word vector extraction method on synonym aggregation. Moreover, the effectiveness of vocabulary translation significantly impacts the quality of the vocabulary list, which challenges the model's robustness.

Acknowledgements

References

Aly Abdelrazek, Walaa Medhat, Eman Gawish, and Ahmed Hassan. 2022. Topic modeling on arabic language dataset: Comparative study. In *International Conference on Model and Data Engineering*, pages 61–71. Springer.

Hervé Abdi. 2010. Coefficient of variation. *Encyclopedia of research design*, 1(5).

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Roman Egger and Joanne Yu. 2022. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7:886498.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Karel Gutiérrez-Batista, Jesús R Campaña, Maria-Amparo Vila, and Maria J Martin-Bautista. 2018. An ontology-based framework for automatic topic detection in multilingual environments. *International Journal of Intelligent Systems*, 33(7):1459–1475.

Johannes Huber and Myra Spiliopoulou. 2019. Learning multilingual topics through aspect extraction from monolingual texts. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 154–183.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Aakanksha Singhal and DK Sharma. 2021. Keyword extraction using renyi entropy: a statistical and domain independent method. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 1970–1975. IEEE.

Louise Smith, Wing Gi Leung, Bryony Crane, Brian Parkinson, Timothea Toulopoulou, and Jenny Yiend. 2018. Bilingual comparison of mandarin and english cognitive bias tasks. *Behavior Research Methods*, 50:302–312.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Nicolas Stefanovitch, Guillaume Jacquet, and Bertrand De Longueville. 2023. Graph and embedding based approach for text clustering: Topic detection in a large multilingual public consultation. In *Companion Proceedings of the ACM Web Conference 2023*, pages 694–700.

Yee Teh, Michael Jordan, Matthew Beal, and David Blei. 2004. Sharing clusters among related groups: Hierarchical dirichlet processes. *Advances in neural information processing systems*, 17.

Silvia Terragni, Elisabetta Fersini, E Fersini, M Passarotti, and V Patti. UMAP2021. Octis 2.0: Optimizing and comparing topic models in italian is even simpler! In *CLiC-it*.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. Octis: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270.

Yu-Xiong Wang and Yu-Jin Zhang. 2012. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on knowledge and data engineering*, 25(6):1336–1353.

Qing Xie, Xinyuan Zhang, Ying Ding, and Min Song. 2020. Monolingual and multilingual topic analysis using lda and bert embeddings. *Journal of Informetrics*, 14(3):101055.

Haiyun Xu, Rui Luo, Jos Winnink, Chao Wang, and Ehsan Elahi. 2022. A methodology for identifying breakthrough topics using structural entropy. *Information Processing & Management*, 59(2):102862.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Zhen Yang, Jianjun Lei, Kefeng Fan, and Yingxu Lai. 2013. Keyword extraction by entropy difference between the intrinsic and extrinsic mode. *Physica A: Statistical Mechanics and its Applications*, 392(19):4523–4531.

A Hyperparameters of Each Topic Model

Model	Hyperparameter	Values/Range
All	Number of topics	[5,20]
LDA	Chunksize	{64,128,......1024}
	Alpha	symmetic,auto
	Eta	symmetic,auto
	Number of passes	[1,20]
	Inerations	{10,30,50,70,90}
	Gamma$_{t}hreshold$	[0.001,0.005]
CTM	Bert model	distiluse-base-multilingual-cased-v2, xlm-mlm-100-1280, xlm-roberta-base,xlm-roberta-large, paraphrase-multilingual-MiniLM-L12-v2, paraphrase-multilingual-mpnet-base-v2
	Activation function	elu,sigmoid,softplus,selu
	Dropout	[0,0.9]
	Number of layers	{1,2,3,4,5}
	Number of samples	{10,50,100,200}
	Momentum	[0,0.9]
	Learning rate	$[10^{-3}, 10^{-1}]$
	Optimizer	adam,sgd
Bertopic	Bert model	aubmindlabbert-large-arabertv2, aubmindlabbert-base-arabertv02-twitter, aubmindlabbert-large-arabertv02, aubmindlabbert-base-arabertv2, xlm-roberta-base, xlm-roberta-large, xlm-mlm-100-1280, paraphrase-multilingual-mpnet-base-v2, distiluse-base-multilingual-cased-v2
HDP	Max_chunks	{10,20,......,50}
	Chunksize	{64,128,256,512,1024}
	Kappa	{-1,-0.8,-0.6,......,1}
NeuralLDA	Dropout	[0,0.9]
	Num_layers	[1,5]
	Batch_size	{64,128,256,512,1024}
NMF	Chunksize	{64,128,256,512,1024}
	Passes	[1,20]
ProdLDA	Dropout	[0,0.9]
	Batch_size	{64,128,256,512,1024}
	Num_layers	[1,5]

Table 3: Hyperparameter settings of topic Models evaluation.

Topic	Zh	Ar
Sports	Quan Hongchan won the title	Cristiano Ronaldo joins Riyadh
warfare	The South Sea dispute of CHN-PH	Israeli-palestinian conflict
Sci. & Tech	Wen Xin Yi Yan release	Reuse of e-waste
economy	Cidic debt	Oil and gold prices rose
political	The twentieth Congress	China-arab summit
Topic	**En**	**Ja**
Sports	Nuggets vs. Lakers in NBA Finals	Shohei Otani won the AL MVP Award
warfare	American troops withdraw from Afghanistan	Situation across the Taiwan Strait
Sci. & Tech	Launch of the Musk Starship	Solar cell "perovskite"
economy	United States inflation	Shibuya large-scale startup support event
political	David Cameron become Britain's foreign secretary	Nuclear sewage discharge
Topic	**Ko**	**Co-concerned**
Sports	2023 Asian Professional Baseball Championship	2022 World Cup Final
warfare	Situation on the Korean Peninsula	Crimean Bridge explosion
Sci. & Tech	Samsung developed its own large model Gauss	Chatgpt4.0 release
economy	Korea's latest GDP release	IMF releases World Economic Outlook
political	APEC meeting	Xi Jinping visit America

Table 4: Five topics and the corresponding events described in each language. The co-concerned section consists of a mix of news under five languages. Each event consists of 20 news items.

Embedding Model (CTM)	Topic Coherence	Topic Diversity
distiluse-base	0.373	0.867
MiniLM	0.468	0.892
MPENT	0.484	0.892
xlm-mlm-100-1280	0.486	0.901
xlm-roberta-base	0.379	0.941
xlm-roberta-large	0.427	0.953

Table 5: Scores of topic coherence and topic diversity with different embedding in CTM.

Embedding Model (BERTopic)	Topic Coherence	Topic Diversity
xlm-roberta-base	0.383	0.873
xlm-roberta-large	0.445	0.907
xlm-mlm-100-1280	0.405	0.823
mpnet	0.515	0.844
MiniLM	0.483	0.925

Table 6: Scores of topic coherence and topic diversity with different embedding in BERTopic.

Topic	Our Model's Output
Topic1:	Qatar, World Cup, football, Argentina, France, Lussel......
Topic2:	Putin, Russia, Ukraine, negotiations, G20, Bahmut, military......
Topic3:	Chatgpt, openai, Intelligence, chat, models, technology, Revolution......
Topic4:	IMF, world, economy, recovery, decline, slowdown, global, US......
Topic5:	Xi,Biden, San Francisco, United States, Fuhrer, arrive......
Topic	BERTopic's Output
Topic1:	game, game, champion, champion, football, history......
Topic2:	war, war, war, tension, situation, atmosphere, troops......
Topic3:	technology, technology, model, model, efficiency, technology......
Topic4:	economy, economy, value, rise, growth, dollar......
Topic5:	meeting, meeting, meeting, meeting, discussion, appointment......

Table 7: The output of our model and BERTopic on the test set, which has been translated into English. Our model outperforms BERTopic in both redundancy and co-concerned topic focus.