RGC Ref. No.:
UGC/FDS16/E01/19
(please insert ref. above)

RESEARCH GRANTS COUNCIL
COMPETITIVE RESEARCH FUNDING SCHEMES FOR
THE LOCAL SELF-FINANCING DEGREE SECTOR

FACULTY DEVELOPMENT SCHEME (FDS)

Completion Report
(for completed projects only)

Submission Deadlines:
1. *Auditor's report with unspent balance, if any: within **six** months of the approved project completion date.*
2. *Completion report: within **12** months of the approved project completion date.*

Part A: The Project and Investigator(s)

1. Project Title

Online Topic Modeling with Applications to Fine-grained Sentiment Analysis

2. Investigator(s) and Academic Department(s) / Unit(s) Involved

Research Team	Name / Post	Unit / Department / Institution
Principal Investigator	WANG Fu-lee, Dean and Professor	School of Science and Technology, Hong Kong Metropolitan University
Co-Investigator	LI Qing, Chair Professor and Head	Department of Computing, The Hong Kong Polytechnic University
Co-Investigator	HAO Tianyong, Professor	School of Computer Science, South China Normal University

3. Project Duration

	Original	Revised	Date of RGC / Institution Approval *(must be quoted)*
Project Start Date	1 January 2020		
Project Completion Date	31 December 2022	31 December 2023	13 October 2022
Duration *(in month)*	36	48	13 October 2022
Deadline for Submission of Completion Report	31 December 2023	31 December 2024	13 October 2022

4.4 Please attach photo(s) of acknowledgement of RGC-funded facilities / equipment.

Part B: The Final Report

5. Project Objectives

5.1 Objectives as per original application

1. To increase the running speed of a topic model over online textual data (i.e., to shorten the convergence time), so that responses can be made in real time for sentiment analysis, information retrieval, question answering, or other online applications.
2. To develop a scalable method of updating the number of topics for our online topic model, in order to capture the burst of new topics and the evolution of existing topics efficiently and effectively.
3. To incorporate the parallel processing of new documents, the online learning of model parameters, and the hierarchical update of topic numbers into fine-grained sentiment analysis systems, so as to use the dynamic topic information to explore the changing rules of user sentiments.

5.2 Revised objectives

Date of approval from the RGC: N/A

Reasons for the change:

1.

2.

3. ....

5.3 Realisation of the objectives

(Maximum 1 page; please state how and to what extent the project objectives have been achieved; give reasons for under-achievements and outline attempts to overcome problems, if any)

In terms of the objectives to be achieved, the research summary of this project is as follows:

(1) To increase the running speed of a flat topic model (i.e., to shorten the convergence time), we developed two efficient and scalable neural topic models, which were published at ***ACL 2020***. On a large corpus, our models can be trained 3 times faster than the existing HDP by GPU parallel computing. For topic hierarchy discovery, we proposed a hierarchical neural topic model to extract a topic tree by exploiting dependency matrices between layers of a network and manifold regularization. Experiments validated that the proposed model outperformed baselines in widely used metrics with much fewer computation costs. The study was published at ***World Wide Web Journal***. For short text modeling, we proposed a context reinforced neural topic model by assuming that each short text covers only a few salient topics. Besides, pre-trained word embeddings were exploited by treating topics as multivariate Gaussian distributions or Gaussian mixture distributions in the embedding space. The study was published at ***Information Sciences***.

(2) To develop a scalable method of updating the number of topics, we employed the finite truncation strategy in our ***ACL 2020*** paper. A truncated model can preserve its nonparametric ability if the number of topics is set to be large enough. To capture the burst of new topics and the evolution of existing topics efficiently and effectively, we proposed a parallel dynamic topic model by developing an adjustment mechanism of evolving topics and reducing the sampling probabilities of topic-indiscriminate words. The model could generate new topics or delete marginal topics according to the evolution of semantics, and improve the quality of topic extraction by a term weighting scheme. The study was published at ***Information Sciences***. To further enhance the topic coherence, we also developed a copula-guided parallel Gibbs sampling algorithm for a model enabling dynamic adjustment of topic numbers, by modeling the topical dependencies within each phrase. The study was published at ***ICDE 2023***.

(3) We exploited document-topic and topic-word distributions generated by a topic model to boost the performance of fine-grained sentiment analysis systems, i.e., emotion distribution learning across domains. Experiments indicate that such topic information was valuable for capturing fine-grained emotions. The study was published at ***Knowledge-Based Systems***. In our study published at ***Information Processing and Management***, we proposed a topic driven adaptive network for cross-domain sentiment classification. To analyze the text of research articles concerning soft computing for sentiment analysis and recommender systems, bibliometrics and structural topic modeling were adopted by our study published at ***IEEE Transactions on Artificial Intelligence***. In our study published at ***Cognitive Computation***, we validated that the local semantics captured by word embedding were also valuable for sentiment analysis. Accordingly, we developed an embedding learning method that can make complete use of the information represented by Chinese characters. Word embedding learning and topic models are complementary in language modeling, which can be trained collaboratively to improve the quality of both word embeddings and topics. In our study published at ***ICTE 2020***, we compared the citation depth for English and Chinese papers. The results indicated several differences which could be categorized by subjects and topics. To integrate the text and image data from social media, we introduced a latent topic memory and proposed a multimodal fusion network for rumor detection. The study was published at ***ICME 2021***. By exploiting a contrary latent topic memory network to store semantic information about true and false patterns of rumors, we also validated that the latent topics were useful for identifying upcoming rumors. The study was published at ***IEEE MultiMedia***. Additionally, neural topic models were beneficial for textual network embedding. In our study published at ***Neural Networks***, we firstly generated document-topic distributions by the neural topic model with Gaussian Softmax constructions. Then, the latent topic relevance information of different vertexes contained in textual attributes information were exploited to enhance the network analysis performance.

5.4 Summary of objectives addressed to date

Objectives *(as per 5.1/5.2 above)*	**Addressed** *(please tick)*	**Percentage Achieved** *(please estimate)*
1. To increase the running speed of a topic model over online textual data (i.e., to shorten the convergence time), so that responses can be made in real time for sentiment analysis, information retrieval, question answering, or other online applications.	✓	100%
2. To develop a scalable method of updating the number of topics for our online topic model, in order to capture the burst of new topics and the evolution of existing topics efficiently and effectively.	✓	100%
3. To incorporate the parallel processing of new documents, the online learning of model parameters, and the hierarchical update of topic numbers into fine-grained sentiment analysis systems, so as to use the dynamic topic information to explore the changing rules of user sentiments.	✓	100%

6. Research Outcome

6.1 Major findings and research outcome

(Maximum 1 page; please make reference to Part C where necessary)

(1) In the field of topic modeling, the rise of neural networks has facilitated the emergence of neural topic models. Given online textual data, we developed two scalable models for dispersed topic discovery, an efficient hierarchical neural topic model with strong interpretability, and a context reinforced neural topic model over short corpora. These studies were published at ***ACL 2020***, ***World Wide Web Journal***, and ***Information Sciences***. Unlike conventional methods, neural topic models can handle large-scale datasets by conveniently exploiting parallel computing facilities like GPUs. The flexibility of neural topic models also enables researchers to tailor model structures to fit textual network embedding and rumor detection. In our study published at ***Neural Networks***, we developed an adversarial capsule network to extract embeddings of textual networks from node structures, vertex attributes, and topics within node text. To ensure a consistent training process by back-propagation, we generated document-topic distributions by the neural topic model with Gaussian Softmax constructions. In our studies published at ***ICME 2021*** and ***IEEE MultiMedia***, we developed two latent topic memory networks to help identify upcoming rumors, by storing semantic information about true and false rumor patterns.

(2) For online topic modeling, it is quite difficult to determine the proper number of topics for a corpus with dynamic document numbers or vocabulary size. Equipped with a monitor-executor mechanism, we first developed a parallel nonparametric topic model that could generate new topics or delete marginal topics according to the evolution of semantics. Then, we proposed a term weighting scheme and a copula guided topical dependency modeling method to improve the topic quality. Both supervised and unsupervised experiments on benchmark datasets validated the efficiency and effectiveness of the proposed model. These studies were published at ***Information Sciences*** and ***ICDE 2023***.

(3) As a hot spot, cross-domain sentiment classification aims to learn a reliable classifier to capture the changing rules of user sentiments. To fully tap the potential of domain-specific information for transfer learning, we first developed a method of extracting domain-specific words based on the topic information derived from topic models. Then, we proposed a semantics attention network and a domain-specific word attention network, the structures of which are based on transformers, for cross-domain sentiment classification. Experiments indicated that topic models are beneficial to sentiment analysis by generating interpretable and low-dimensional subspaces. The study was published at ***Information Processing and Management***.

(4) Emotion distribution learning, which aims to predict the intensity values of an instance over a set of emotion categories, is one of the main tasks in fine-grained sentiment analysis. However, the previous methods had limited performance when the domain of testing instances differed from that of the training set. In our study published at ***Knowledge-Based Systems***, we proposed a constrained optimization approach based on non-negative matrix tri-factorization (NMTF) for cross-domain emotion distribution learning. Experiments indicated that the model performance was both boosted and stabilized when the output of a topic model was employed to initialize the key matrices in NMTF.

(5) The topics extracted from topic models are also valuable for downstream tasks such as content analysis. In our study published at ***IEEE Transactions on Artificial Intelligence***, we adopted bibliometrics and structural topic modeling to analyze the text contents of research articles concerning soft computing for sentiment analysis and recommender systems. Furthermore, we explored the association between topic modeling and representation (or embedding) learning. In our study published at ***Cognitive Computation***, we developed an embedding learning method that can make complete use of the information represented by Chinese characters, including phonology, morphology, and semantics. In our study published at ***ACM Transactions on Intelligent Systems and Technology***, we summarized and categorized the contrastive learning based sentence representation models, which shed light on sentence classification, sentiment analysis, machine translation, question answering, and many other downstream tasks.

6.2 Potential for further development of the research and the proposed course of action *(Maximum half a page)*

Recently, embedding (or representation) learning and topic modeling have been highlighted by researchers as complementary factors. These two tasks can be trained collaboratively to improve the quality of both embeddings and topics. As a novel paradigm for representation learning and topic modeling over text, non-negative matrix tri-factorization (NMTF) has attracted much attention. However, NMTF involves intensive matrix multiplications. To address this, we plan to develop a distributed-memory parallel algorithm to accelerate NMTF on large-scale text. The above method will be conducted in parallel without sacrificing model quality. Furthermore, text content analysis has been considered by many researchers as the dominant use case for topic modeling. In our project, we have adopted bibliometrics and structural topic modeling to analyze the text contents of research articles concerning soft computing for sentiment analysis and recommender systems. The study was published at ***IEEE Transactions on Artificial Intelligence***. To investigate the possibility of automatic classification for the semantic content of the possibility of automatic classification for the semantic content of MOOC course reviews, we have also discovered seven course factors that learners frequently mentioned by summarizing 186,738 review sentences. Subsequently, each factor was assigned a sentimental value and the topics that could influence learners' learning experiences the most were decided. The study was published at ***Future Internet***. Considering the important role of topic models in text content analysis, we plan to identify the relationships between per-document covariates (e.g., source, author, and style) and the discovered topics, so as to facilitate the prediction of new documents with unseen covariate configurations and better understanding of each topic.

7. Layman's Summary

(Describe <u>in layman's language</u> the nature, significance and value of the research project, in no more than 200 words)

Topic modeling can automatically reveal the semantic structure in text and deal with the ambiguity caused by synonymous and polysemous words. Thus, it has been widely used in natural language processing. However, determining the proper number of topics for a corpus is difficult — with too small number of topics, the global semantic structure may not be well discovered, but with too large number of topics, the semantics of topics may be overlapped with each other. The hierarchical Dirichlet process (HDP) is a well-known approach to automatically find an optimal number of topics, but it requires massive computational resources for parameter estimation. In this project, we have developed scalable methods of updating the number of topics, neural topic models that could conveniently exploit parallel computing facilities, tailored models of using topics for fine-grained sentiment analysis, textual network embedding, and rumor detection. Relevant research results were published at reputable journals and conferences, including ACM Transactions on Intelligent Systems and Technology, IEEE Transactions on Artificial Intelligence, IEEE MultiMedia, Neural Networks, Information Processing and Management, World Wide Web Journal, Knowledge-Based Systems, Information Sciences, ACL 2020, ICME 2021, and ICDE 2023.

Part C: Research Output

8. Peer-Reviewed Journal Publication(s) Arising Directly From This Research Project

(Please attach a copy of the publication and/or the letter of acceptance if not yet submitted in the previous progress report(s). All listed publications must acknowledge RGC's funding support by quoting the specific grant reference.)

The Latest Status of Publications				Author(s) *(denote the corresponding author with an asterisk*)*	Title and Journal / Book *(with the volume, pages and other necessary publishing details specified)*	Submitted to RGC *(indicate the year ending of the relevant progress report)*	Attached to this Report *(Yes or No)*	Acknowledged the Support of RGC *(Yes or No)*	Accessible from the Institutional Repository *(Yes or No)*
Year of Publication	**Year of Acceptance** *(For paper accepted but not yet published)*	**Under Review**	**Under Preparation** *(optional)*						
2021				Qinjuan Yang, Haoran Xie, Gary Cheng*, Fu Lee Wang, Yanghui Rao	Pronunciation-Enhanced Chinese Word Embedding. *Cognitive Computation*, 13, 688-697. https://doi.org/10.1007/s12559-021-09850-9	Yes 2021	No	Yes	Yes
2021				Xiaorui Qin, Yufu Chen, Yanghui Rao, Haorao Xie*, Man Leung Wong, Fu Lee Wang	A Constrained Optimization Approach for Cross-Domain Emotion Distribution Learning, *Knowledge-Based System*, 227, 107160. https://doi.org/10.1016/j.knosys.2021.107160	Yes 2021	No	Yes	Yes
2021				Ziye Chen, Cheng Ding, Yanghui Rao*, Haoran Xie, Xiaohui Tao, Gary Cheng, Fu Lee Wang	Hierarchical neural topic modeling with manifold regularization, *World Wide Web*, 24(6), 2139-2160. https://doi.org/10.1007/s11280-021-00963-7	Yes 2022	No	Yes	Yes
2021				Xiaorui Qin, Yanghui Rao*, Haoran Xie, Jiahai Wang, Fu Lee Wang	TACN: A Topical Adversarial Capsule Network for textual network embedding, *Neural Networks*, 144, 766-777. https://doi.org/10.1016/j.neunet.2021.09.026	Yes 2022	No	Yes	Yes

2022				Xieling Chen, Fu Lee Wang*, Gary Cheng, Man-Kong Chow, Haoran Xie	Understanding learners' perception of MOOCs based on review data analysis using deep learning and sentiment analysis, *Future Internet*, 14(8), 218. https://doi.org/10.3390/fi14080218	Yes 2022	No	Yes	Yes
2022				Jiaxin Chen, Zekai Wu, Zhenguo Yang*, Haoran Xie, Fu Lee Wang, Wenyin Liu	Multimodal Fusion Network With Contrary Latent Topic Memory for Rumor Detection, *IEEE MultiMedia*, 29, 1, 104-113, https://doi.org/10.1109/MMUL.2022.3146568	Yes 2022	No	Yes	Yes
2022				Jiachun Feng, Zusheng Zhang, Cheng Ding, Yanghui Rao*, Haoran Xie, Fu Lee Wang	Context reinforced neural topic modeling over short texts, *Information Sciences*, 607, 79-91, https://doi.org/10.1016/j.ins.2022.05.098	Yes 2022	No	Yes	Yes
2022				Hongyu Jiang, Zhiqi Lei, Yanghui Rao*, Haoran Xie, Fu Lee Wang,	Parallel dynamic topic modeling via evolving topic adjustment and term weighting scheme, *Information Sciences*, 585, 176-193, https://doi.org/10.1016/j.ins.2021.11.060	Yes 2022	No	Yes	Yes
2022				Xieling Chen, Haoran Xie*, Jingjing Wang, Zongxi Li, Gary Cheng, Man Leung Wong, Fu Lee Wang	A Bibliometric Review of Soft Computing for Recommender Systems and Sentiment Analysis, *IEEE Transactions on Artificial Intelligence*, 3, 5, 642-656, https://doi.org/10.1109/TAI.2021.3116551	No Att. 1	Yes	Yes	Yes

2022				Fu Lee Wang, Yuyin Lu, Gary Cheng*, Haoran Xie, Yanghui Rao	Learning Chinese word embeddings from semantic and phonetic components, *Multimedia Tools and Applications*, 81, 42805–42820, https://doi.org/10.1007/s11042-022-13488-6	No Att. 2	Yes	Yes	Yes
2023				Xieling Chen, Haoran Xie*, Zongxi Li, Dian Zhang, Gary Cheng, Fu Lee Wang, Hong-Ning Dai, Qing Li	Leveraging deep learning for automatic literature screening in intelligent bibliometrics, *International Journal of Machine Learning and Cybernetics*, 14, 1483–1525 https://doi.org/10.1007/s13042-022-01710-8	No Att. 3	Yes	Yes	Yes
2023				Fu Lee Wang, Zhengwei Zhao, Gary Cheng*, Yanghui Rao, Haoran Xie	Weighted cluster-level social emotion classification across domains, *International Journal of Machine Learning and Cybernetics*, 14, 2385–2394 https://doi.org/10.1007/s13042-022-01769-3	No Att. 4	Yes	Yes	Yes
2023				Jingjing Wang, Haoran Xie, Fu Lee Wang*, Lap-Kei Lee	Improving text classification via a soft dynamical label strategy, *International Journal of Machine Learning and Cybernetics*, 14, 2395–2405 https://doi.org/10.1007/s13042-022-01770-w	No Att. 5	Yes	Yes	Yes
2023				Zongxi Li, Xianming Li, Haoran Xie, Fu Lee Wang*, Mingming Leng, Qing Li, Xiaohui Tao	A novel dropout mechanism with label extension schema toward text emotion classification, *Information Processing and Management*, 60, 2, 103173, https://doi.org/10.1016/j.ipm.2022.103173	No Att. 6	Yes	Yes	Yes

2023				Jingjing Wang, Haoran Xie, Fu Lee Wang*, Lap-Kei Lee, Mingqiang Wei	Jointly modeling intra- and inter-session dependencies with graph neural networks for session-based recommendations, *Information Processing and Management*, 60, 2, 103209, https://doi.org/10.1016/j.ipm.2022.103209	No Att. 7	Yes	Yes	Yes
2023				Yicheng Zhu, Yiqiao Qiu, Qingyuan Wu*, Fu Lee Wang Yanghui Rao	Topic Driven Adaptive Network for cross-domain sentiment classification *Information Processing and Management*, 60, 2, 103230, https://doi.org/10.1016/j.ipm.2022.103230	No Att. 8	Yes	Yes	Yes
2023				Jingjing Wang, Haoran Xie, Fu Lee Wang*, Lap-Kei Lee	A transformer–convolution model for enhanced session-based Recommendation, *Neurocomputing*, 531, 21–33, https://doi.org/10.1016/j.neucom.2023.01.083	No Att. 9	Yes	Yes	Yes
2023				Zhongyu Zhuang, Ziran Liang, Yanghui Rao*, Haoran Xie, Fu Lee Wang	Out-of-vocabulary word embedding learning based on reading comprehension mechanism, *Natural Language Processing Journal*, 5, 100038, https://doi.org/10.1016/j.nlp.2023.100038	No Att. 10	Yes	Yes	Yes
2023				Xinhong Chen, Qing Li, Zongxi Li*, Haoran Xie*, Fu Lee Wang, Jianping Wang	A Reinforcement Learning Based Two-Stage Model for Emotion Cause Pair Extraction, *IEEE Transactions on Affective Computing*, 14, 3, 1779-1790, https://doi.org/10.1109/TAFFC.2022.3218648	No Att. 11	Yes	Yes	Yes

2023				Lingling Xu, Haoran Xie*, Zongxi Li, Fu Lee Wang, Weiming Wang, Qing Li	Contrastive Learning Models for Sentence Representations, *ACM Transactions on Intelligent Systems and Technology*, 14, 4, 67, *https://doi.org/10.1145/3593590*	No Att. 12	Yes	Yes	Yes
2024				Pengbo Mao, Hegang Chen, Yanghui Rao*, Haoran Xie, Fu Lee Wang	Contrastive learning for hierarchical topic modelling, *Natural Language Processing Journal*, 6, 100058. https://doi.org/10.1016/j.nlp.2024.100058	No Att. 13	Yes	Yes	Yes

9. Recognized International Conference(s) In Which Paper(s) Related To This Research Project Was / Were Delivered

(Please attach a copy of each conference abstract)

Month / Year / Place	**Title**	**Conference Name**	**Submitted to RGC** *(indicate the year ending of the relevant progress report)*	**Attached to this Report** *(Yes or No)*	**Acknowledged the Support of RGC** *(Yes or No)*	**Accessible from the Institutional Repository** *(Yes or No)*
July 2020, Online	Neural Mixed Counting Models for Dispersed Topic Discovery	The 58th Annual Meeting of the Association for Computational Linguistics (ACL)	Yes 2021	No	Yes	Yes
August 2020, Online	Influence Evaluation of Academic Papers via Citation Characteristics Analysis	The 5th International Conference on Technology in Education (ICTE)	Yes 2022	No	Yes	Yes
July 2021, Shenzhen, China	Multimodal Fusion Network with Latent Topic Memory for Rumor Detection	IEEE International Conference on Multimedia and Expo (ICME)	Yes 2022	No	Yes	Yes
April 2023, Anaheim, CA, USA	Copula Guided Parallel Gibbs Sampling for Nonparametric and Coherent Topic Discovery (Extended Abstract)	2023 IEEE 39th International Conference on Data Engineering (ICDE)	No	Yes Att. 14	Yes	Yes

10. Whether Research Experience And New Knowledge Has Been Transferred / Has Contributed To Teaching And Learning

(Please elaborate)

One PhD student was involved in the project. This project contributed to training of research student.

11. Student(s) Trained

(Please attach a copy of the title page of the thesis)

Name	Degree Registered for	Date of Registration	Date of Thesis Submission / Graduation
Nil			

12. Other Impact

(e.g. award of patents or prizes, collaboration with other research institutions, technology transfer, teaching enhancement, etc.)

We have established collaboration with research partners from a number of universities in Hong Kong and Chinese Mainland.

13. Statistics on Research Outputs

	Peer-reviewed Journal Publications	Conference Papers	Scholarly Books, Monographs and Chapters	Patents Awarded	Other Research Outputs (please specify)	
No. of outputs arising directly from this research project	21	4	0	0	Type	No.

14. Public Access Of Completion Report

(Please specify the information, if any, that cannot be provided for public access and give the reasons.)

Information that Cannot Be Provided for Public Access	Reasons