RGC Ref. No.:
UGC/FDS11/E03/16
(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

Sentiment Analysis based on Multi-source Social Network Data

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

Research Team	Name / Post	Unit / Department / Institution
Principal Investigator	WANG Fu-lee / Dean (01/01/2017 – 05/03/2018)	School of Science and Technology / Open University of Hong Kong
	Prof POON CHUNG-keung / Director (05/03/2018 – 05/06/2018) RGC Approval: 05/03/2018	School of Computing and Information Sciences / Caritas Institute of Higher Education
	Dr ZHAO Yingchao / Associate Professor (05/06/2018 – 31/12/2019) RGC Approval: 05/06/2018	School of Computing and Information Sciences / Caritas Institute of Higher Education
Co-Investigator(s)	WANG Fu-lee / Dean	School of Science and Technology / Open University of Hong Kong
Co-Investigator(s)	XIE Haoran / Associate Professor	Department of Computing and Decision Sciences / Lingnan University
Co-Investigator(s)	LIU Wenyin / Professor	School of Computer Science / Guangdong University of Technology

3. Project Duration

	Original	**Revised**	**Date of RGC / Institution Approval** *(must be quoted)*
Project Start Date	01/01/2017	N/A	N/A
Project Completion Date	31/12/2019	N/A	N/A
Duration *(in month)*	36	N/A	N/A
Deadline for Submission of Completion Report	31/12/2020	N/A	N/A

Part B: The Final Report

5. Project Objectives

5.1 Objectives as per original application

1. Expand short documents in social networks from the whole corpus. In this task, we aim to incorporate the most similar words of related documents into each short document, so as to alleviate the issue of data sparsity and diversity for multi-source social network text.
2. Develop de-noising models to identify noisy sentimental labels within cross-domain social network data. One objective of this task is to infer the actual sentimental category of each document from the noisy annotated text, so as to improve the performance of sentiment classifiers. The other purpose is to identify noisy annotators by our models, which can be applied to opinion spam detection and quality control in crowdsourcing.
3. Extract sentiments from multi-source social network data comprehensively and accurately by jointly modeling the users, text and labels in the latent topic model. In this task, we aim to utilize multiple dimensions in social networks simultaneously, so as to improve the effectiveness of sentiment analysis models.
4. Propose an event-based framework of dynamic sentiment analysis. This task is to conduct dynamic sentiment analysis on various social network data via a high-efficiency strategy.
5. Apply sentiment analysis to other related domains. The goal of this task is to employ methods and techniques of sentiment analysis in some domain-specific applications like stock price prediction, e-learning promotion and personalized search.

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)

All the objectives proposed are realized with the extent of each explained as follows:

1. **Expand short documents in social networks from the whole corpus.**
   We proposed the concept of sentimental context, and explored external sentiment lexicons to extract the sentimental context. By integrating the sentimental context with either a word-level or a topic-level sentiment classification model, the performance of sentiment classification over short text was improved [C2]. To accurately detect social emotions over short messages from the perspective of readers in social network, a universal affective model was further developed in [J2]. We extracted biterms to enrich features of a short text, where a biterm denotes an unordered word-pair co-occurring in a proper text window of the document. Then we combined biterms with keywords extracting by a new paradigm named Average Term Frequency Inverse Document Frequency (ATF-IDF) to enhance the semantic relationships between biterms.
2. **Develop de-noising models to identify noisy sentimental labels within cross-domain social network data.**
   We proposed a probabilistic de-noising model to infer the actual sentimental category of each document from the noisy annotated text and to identify noisy users [C1]. In online review analysis, we observed that the sentiment of a document might be combined or more fine-grained than a single positive or negative polarity. Therefore, we proposed a segment-level joint topic-sentiment model (STSM) by incorporating conjunctions and prior knowledge from sentiment lexicons into the training process [J3]. The experimental results on sentiment classification indicated that the proposed method could boost performance for compound and complex sentences.
3. **Extract sentiments from multi-source social network data comprehensively and accurately by jointly modeling the users, text and labels in the latent topic model.**
   We proposed a supervised user-product topic model in [C4] to incorporate user and product topic preferences into non-negative matrix tri-factorization to factorize ratings. Experiments on two popular datasets which contain users' reviews and sentimental ratings indicated that our model can outperform baselines consistently. To explore fine-grained associations between text and labels, a Siamese network-based supervised topic model was further developed in [C6]. Since topics play an important role in understanding sentences or user comments, the model performance was evaluated by predicting emotions and sentiment strengths over multi-source social network data. This model could also effectively predict categorical or real-valued labels for new documents by generating word embedding from a label-specific topical space.
4. **Propose an event-based framework of dynamic sentiment analysis.**
   We proposed a novel framework, which uses the emotion distribution of training documents at the cluster level, to conduct emotion detection dynamically and efficiently in [C3]. We developed a novel model of semantically rich hybrid neural network (HNN) which leverages unsupervised teaching models to incorporate semantic domain knowledge into the neural network to bootstrap its inference power and interpretability [J1].
5. **Apply sentiment analysis to other related domains.**
   We proposed methods based on recurrent neural network and convolutional neural network to detect rumors automatically in social media [C8]. To explore the connection between sentiment classification and semantic composition, we employed several sentiment classification models to capture typical semantic composition cases, including negation reversing (e.g., not interesting), negation shifting (e.g., not terrific), and intensification (e.g., very good). We can use these methods for movie reviews [C5] [J4]. We developed extractive adversarial networks to extract the latent space from labels, attributed information, and topological structure jointly [C7].

5.4 Summary of objectives addressed to date

Objectives *(as per 5.1/5.2 above)*	**Addressed** *(please tick)*	**Percentage Achieved** *(please estimate)*
1. Expand short documents in social networks from the whole corpus. In this task, we aim to incorporate the most similar words of related documents into each short document, so as to alleviate the issue of data sparsity and diversity for multi-source social network text.	√	100%
2. Develop de-noising models to identify noisy sentimental labels within cross-domain social network data. One objective of this task is to infer the actual sentimental category of each document from the noisy annotated text, so as to improve the performance of sentiment classifiers. The other purpose is to identify noisy annotators by our models, which can be applied to opinion spam detection and quality control in crowdsourcing.	√	100%
3. Extract sentiments from multi-source social network data comprehensively and accurately by jointly modeling the users, text and labels in the latent topic model. In this task, we aim to utilize multiple dimensions in social networks simultaneously, so as to improve the effectiveness of sentiment analysis models.	√	100%
4. Propose an event-based framework of dynamic sentiment analysis. This task is to conduct dynamic sentiment analysis on various social network data via a high-efficiency strategy.	√	100%
5. Apply sentiment analysis to other related domains. The goal of this task is to employ methods and techniques of sentiment analysis in some domain-specific applications like stock price prediction, e-learning promotion and personalized search.	√	100%

6. Research Outcome

6.1 Major findings and research outcome
(Maximum 1 page; please make reference to Part C where necessary)

In the study of Objective 1, we proposed the concept of sentimental context to enrich the characteristics and improved the performance of sentiment classification over short text and integrated it with sentiment classifiers in [C2]. To address the issue of limited words in each short message, we further proposed a universal affective model (UMA) for social emotion detection by exploiting keywords, biterms, and background words jointly in [J2]. On the task of short text emotion classification, experimental results indicated that the proposed UAM outperformed state-of-the-art deep learning models in most cases.

In the study of Objective 2, we proposed a network framework in [C1] to aggregate noisy labels and identify malicious users simultaneously. A new simulation approach was also developed to generate noisy annotations by following the process of Profile Injection Attack in Collaborative Recommender Systems. Since it may be rough to use labels as the only supervision information to train a sentiment classifier, we further developed a segment-level joint topic-sentiment model for online review analysis in [J3]. Experimental results indicated that it was beneficial to sentiment classification by splitting each sentence into segments using conjunctions and exploiting sentiment lexicons.

In the study of Objective 1, to utilize the multi-dimensional features effectively, we proposed a supervised user-product topic model [C4] by capturing both user preferences and attractive characteristics of products. Two benchmark datasets which contain users' reviews and ratings on different aspects of movies and on different restaurants were employed for evaluation. Experimental results indicated that the recommendation performance could be enhanced by modeling the users, text, and labels jointly. Since topics are critical to sentiment analysis, we further developed a Siamese network-based supervised topic model [C6] which can jointly model documents and labels. Extensive experiments validated that the proposed model could generate accurate label-specific topics for sentiment and emotion detection.

In the study of Objective 4, to capture the dynamic of sentiments and emotions, we proposed a cluster-level classifier with an event-based training process in [C3]. A new dataset with multi-domain news articles and reader ratings over 8 emotions was collected for evaluation. Experimental results indicated that the proposed cluster-level classifier outperformed the existing word-level and topic-level models on identifying varied and evolved reader emotions toward news events across domains. Furthermore, a high-efficiency hybrid neural network was developed to enhance the inference power and interpretability of social emotion classification models [J1]. Based on three real-world social network datasets, experimental results confirmed that the proposed model outperformed conventional neural networks and other state-of-the-art methods.

Finally, in the study of Objective 5, we applied sentiment analysis methods and techniques to other domain-specific applications, such as rumor detection in social network [C8]. The research results on the connection between sentiment classification and semantic composition were shown in [C5] and [J4]. We also presented a network embedding method for multi-source social network data analysis in [C7].

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

Our sentiment analysis models are used to analyse a large amount of data in social networks, which usually requires long time to do data training and testing. One research direction is to reduce the complexity of our models so as to accelerate the data training and testing.

We will first consider to use parallel techniques. There are two types of parallel architectures: multi-processor architecture (MPA) and multi-core architecture (MCA). Since the communication cost of a typical sentiment analysis model is often greater than the computation cost with the increase of the number of processors, which leads to serious scalability issues, it is better to use multi-core architecture (MCA), where all threads share the same memory space. However, it is critical to achieve the data synchronization in MCA. Another possible way is to exploit sparse coding of clustering algorithms.

7. Layman's Summary

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

The research project concerns the sentiment analysis problem. The ultimate goal is to automatically extract sentiment-related information from text in multi-source social network. Nowadays, more and more people express their opinions on the Internet. Short texts (e.g., microblog, brief comments), which contain users' opinions, are becoming increasingly prevalent. These short texts can express writers' feeling and evoke readers' emotions. However, sentiment analysis on short texts are quite different from traditionally sentiment analysis on long articles because the pieces of information in social network are usually very short, user relevant, and containing lots of noisy labels.

With the techniques proposed in this project, we can expand the short texts, identify noisy sentimental labels, and extract sentiments from social networks comprehensively and accurately by jointly modeling the users, text and labels in the latent topic model. The research results have not only addressed the challenging issues triggered by data sparsity, noisy labels, varied user emotion perception, and sentiment evolution in multi-source social network data, but also shed light on computational social science, rumor detection, network embedding, and other areas.

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)*						
2017				Xiangsheng Li, Yanghui Rao*, Haoran Xie, Raymond Y. K. Lau, Jian Yin, Fu Lee Wang	[J1] Bootstrapping Social Emotion Classification with Semantically Rich Hybrid Neural Networks. IEEE Transactions on Affective Computing. 8(4): 428-442	No	Yes	Yes	**Yes**
2018				Weiming Liang, Haoran Xie*, Yanghui Rao, Raymond Y. K. Lau, Fu Lee Wang	[J2] Universal Affective Model for Readers' Emotion Classification over Short Texts. Expert Systems With Applications. 114: 322-333	No	Yes	Yes	**Yes**
2019				Qinjuan Yang, Yanghui Rao, Haoran Xie, Jiahai Wang, Fu Lee Wang, Wai Hong Chan	[J3] Segment-Level Joint Topic-Sentiment Model for Online Review Analysis. IEEE Intelligent Systems. 34(1): 43-50	No	Yes	Yes	**Yes**

2019				Xingming Chen, Yanghui Rao*, Haoran Xie, Fu Lee Wang, Yingchao Zhao, Jian Yin	[J4] Sentiment Classification Using Negative and Intensive Sentiment Supplement Information. Data Science and Engineering. 4(2): 109-118	No	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)*
07-08 / 2017 / Vancouver, Canada	A Network Framework for Noisy Label Aggregation in Social Media	[C1]The 55th Annual Meeting of the Association for Computational Linguistics (ACL)	Yes (2018)	Yes	Yes	Yes
10 / 2017 / Krakow, Poland	Sentiment Classification of Short Text Using Sentimental Context	[C2]The 2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC)	Yes (2018)	Yes	Yes	Yes
11 / 2017 / Singapore	Cluster-level Emotion Pattern Matching for Cross-Domain Social Emotion Classification	[C3]The 26th ACM International Conference on Information and Knowledge Management (CIKM)	Yes (2018)	Yes	Yes	Yes
05 / 2018 / Gold Coast, QLD, Australia	Learning Dual Preferences with Non-negative Matrix Tri-Factorization for Top-N Recommender System	[C4]The 23rd International Conference on Database Systems for Advanced Applications (DASFAA)	Yes (2018)	Yes	Yes	Yes

07 / 2018 / Macau, China	Sentiment Classification via Supplementary Information Modeling	[C5]The 2nd International Joint Conference on Web and Big Data (APWeb-WAIM)	No	Yes	Yes	Yes
10-11 / 2018 / Brussels, Belgium	Siamese Network-Based Supervised Topic Modeling	[C6]The 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)	No	Yes	Yes	Yes
11 / 2018 / Kaohsiung, Taiwan	Extractive Adversarial Networks for Network Embedding	[C7]The 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC)	No	Yes	Yes	Yes
06 / 2019 / Daejeon, South Korea	Supervised Group Embedding for Rumor Detection in Social Media	[C8]The 19th International Conference on Web Engineering (ICWE)	No	Yes	Yes	Yes

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

(Please elaborate)

The research outputs are novel techniques to demonstrate the application of Natural Language Processing. The knowledge from this project has been added to the course of Artificial Intelligence within our school's artificial intelligence degree programme to enrich students' knowledge.

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.)

Nil

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	4	8			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
N/A	