Discovery of learning topics in an online social network for health professionals

Xin LI[a,1], Karin VERSPOOR[a], Kathleen GRAY[a] and Stephen BARNETT[b]
[a] *Health and Biomedical Informatics Research Centre, University of Melbourne*
[b] *General Practice Academic Unit, Graduate School of Medicine, University of Wollongong*

Abstract. Online social networking provides support to health professionals' learning and professional development. To understand their learning needs in this context, this study employs topic modelling of postings to an online social network for health professionals to identify the topics of interest. The analysis shows that the health professionals in this network were more interested in discussing non-clinical topics than clinical ones. The non-clinical topics include some controversial topics such as policy-related issues, as well as an interest in the latest news and advanced information in the field. The clinical topics relate to their practices, including sharing practical and experiential knowledge and providing benchmarks.

Keywords. Topic modelling, networked learning, health professional education

1. Introduction

As medical knowledge expands and healthcare delivery becomes more complex, health professionals must commit to continuous learning to maintain up-to-date knowledge and skills. One approach to meeting their learning and development needs is through engagement in an online social network (OSN) [1]. OSNs have been found useful to reduce professional isolation and support anytime-anywhere peer-to-peer interaction at scale. Also, they are thought to contribute to the development of professional networks and improve continuing professional development.

There are many OSN targeted towards health professionals but they appear to fail to support the broader learning objectives [2]. It has been recognised that there is a lack of understanding about how health professionals learn in an OSN, making it difficult to design and facilitate this type of learning [3]. To realise the full potential of OSNs for health professionals' learning, understanding and evaluating this learning context is important.

Previous studies focused on understanding learning behaviours by identifying the patterns of the interaction among health professionals [4, 5]. However, there is still much to be explored in terms of the textual dialogue among health professionals, particularly regarding how those dialogues support the process of learning. This paper proposes topic modelling as a method to discover the topics of interest from an OSN for health professionals. The identified topics can provide insights on the learning resource and professional development needs of the health professionals.

2. Background and Related Work

Previous work has been done on analysis of dialogue in online learning environments to find evidence about learning and knowledge construction. This has required considerable resources and effort for manual data coding to analyse cognitive and social processes in which learners engage. For example, De Laat [6] assessed the quality of the dialogue in an online community for the police using a coding scheme that examines the social construction of knowledge. Schrire [7] investigated the

[1] Corresponding Author: Xin Li; E-mail: xinli87@gmail.com.

knowledge-building process in a discussion forum used in a higher education context by applying community of inquiry model.

As more and more textual data is generated online and human annotation becomes impossible, computational tools such as topic modelling become more useful. Topic modelling is a statistical method that analyses the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time [8].

Many researchers have used topic modelling to explore the themes in dialogues in online learning environments. However, to the best of our knowledge, its application in discovering topics among the online community of health professionals is novel in health professional education research. Tobarra, Robles-Gómez [9] used it to discover topics of interest in the forum of a Learning Management System for improving the structure and contents of education courses. Portier, Greer [10] used it together with sentiment analysis to identify improvements that enhance social support in an online cancer community. Most recently, Ezen-Can, Boyer [11] used it to understand the topics of discussion in the forum of an open online course for educators.

3. Methods

3.1. Dataset

Data were collected from the database of an online discussion forum provided by a health professional OSN host organisation, with Human Research Ethics approval. The online forum was established in 2009 specifically for registered health practitioners and had more than 10,000 members. Since the online forum was set up for doctors to discuss industry issues, share best practices and promote conversation within the health community, it is logical to assume that the topics discovered from the forum posts would reflect the resource and professional development needs of this community.

The data for this study comprised all the posts made by the forum participants (N = 48) who remained active in three consecutive years from the period 2012 to 2014. The three-year period represents 50% of the overall operating period of this forum, and the most recent and complete years available at the time of data collection in 2015. The 48 forum participants represent 13% of overall participants during this period. 154 discussion threads were found, each receiving between one and 58 replies. A total of 1604 posts (105,063 words) were extracted from the forum.

3.2. Topic Modeling Using MALLET

To identify the topics of interest in this forum, we generated a topic model using the MALLET tool implemented in R. The MALLET (Machine Learning for Language Toolkit) automates the process of topic discovery from a large volume of text; it implements the latent Dirichlet allocation (LDA) algorithm, which is a generative probabilistic model [12]. The basic idea of LDA is that documents are presented as a random mixture of topics, where each topic is a probability distribution over a given vocabulary of words [13]. In this study, a document is defined as a forum post. The MALLET program was used to generate clusters of words (i.e. topics) that frequently occur together within a forum post.

3.3. Procedure

Data preparation: We pulled full text from each post using SQL queries, and cleaned the text by removing anything other than English letters or spaces. To improve the coherence of generated topics, we removed the stop words from the full text based on the standard list of stop words of MALLET[2]. We also further removed popular words (e.g. lol, cheers, pretty, nice, yrs) and any specific words associated with country/state/city and personal names (e.g. sherlock, watson, judas) that appeared in

[2] For instructions to download the standard list of stop words of MALLET, please go to http://mallet.cs.umass.edu/import-stoplist.php.

this dataset. In addition, all words were stemmed to retrieve their stems so that various forms of a word would be counted together when counting word frequency. This was done using the stemmer function (available in `tm` package) in `R`. These pre-processing steps reduced the number of words in the dataset to 54873.

Topic model generation: To generate topic models using MALLET, two variables (i.e. number of topics, number of sampling iterations) must be defined. To identify the optimal number of topics for the topic model, we specified different numbers of topics to generate four models (Models 1 – 4). The initial number of topics was set to 15 by inspecting all the 154 thread titles and noting from inspection that there are approximately 14 broad topics in the dataset. The dataset of this size usually has the default sampling iteration set to 400. Since increasing the number of iterations may improve topic coherence [12], we increased the iteration to 800 to generate two additional models (Models 5 – 6). Table 1 depicts the variables defined for these different topic models.

Table 1. The variables of various topic models

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Number of Topics T	15	20	25	30	20	25
Number of Iterations I	400	400	400	400	800	800

Topic inference: Topics were inferred using clusters of words produced by topic models. Since each topic is a probability distribution over words, we chose to inspect the top ten words for each cluster. This is based on the assumption that more words per cluster might make it more difficult to infer a meaningful topic for each cluster.

Topic optimisation: Inferred topics were optimised by reviewing the contents of the top five posts with consideration for each word cluster (i.e. topic). The top five posts for each topic were identified by inspecting the probability of each topic appearing in each post, which was obtained by employing the function `mallet.doc.topics` in MALLET. The optimisation helped identify further duplicates and improve the accuracy of the inferred topics.

Topic evaluation: Traditionally, the performance of topic models are typically evaluated using quantitative intrinsic methods such as computing the probability of held-out documents. However, it has been shown that this measure is not always a good predictor of human judgment [14]. In this study, we evaluated the topics based on human judgment using F-measure [15], which is often used in the field of information retrieval. There are four performance metrics considered (i.e. accuracy, precision, recall, and F-score, as defined in Table 2).

Table 2. The performance metrics

Performance metric	**Description**	**Formula**
Accuracy	The percentage of the posts identified are expected to belong to an optimised topic.	TP + TN / (TP + FP + FN + TN)
Precision	The percentage of posts correctly identified as belonging to an optimised topic.	TP / (TP + FP)
Recall	The percentage of posts identified as belonging to any topic.	TP / (TP + FN)
F-score	The harmonic mean of precision and recall, which can be interpreted as a weighted average of the Precision and Recall.	2 × (Precision × Recall) / (Precision + Recall)

We randomly selected 40 forum posts from the dataset to validate the optimised topics using F-measure. We considered that a post is True Positive (TP) when any part of the post content matches an optimised topic; a post is False Positive (FP) when the post content does not match an optimised topic; a post is False Negative (FN) when the post content suggests a discernable topic (may be an optimised topic or any new topic that has not been identified); a post is True Negative (TN) when the post content does not suggest any discernable topic.

4. Results and Discussion

4.1. Topic Model Comparison

After inferring the topics of generated topic models, we compared the number of optimised topics from each topic model. It seems that a topic model with T = 20 would be more appropriate than T = 15, or T = 25, or T = 30. As shown in Table 3, Model 1 (T = 15) generated 9 optimised topics which indicate that setting too few topics could result in not covering all topics. Model 3 (T = 25) generated 11 optimised topics which indicate that setting too many topics could result in duplications (five pairs of word clusters represent the same topic). Model 4 (T = 30) generated only 9 optimised topics which indicate that setting too many topics could even result in uninterpretable topics.

For this dataset, a topic model with I = 400 would be more appropriate than I = 800. The number of optimised topics generated from Model 5 and Model 6 suggests that increasing the number of iterations did not result in better topic models, as the composition and quality of the resulting topics only increased to a certain point and then levelled off. From the results, we concluded that Model 2 (T = 20, I = 400) seems to be the topic model that best describes the topics of interest discussed by health professionals in this forum.

Table 3. Number of inferred and optimised topics for various topic models

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Number of inferred topics	11	14	15	13	13	12
Number of optimised topics	9	13	11	9	12	9

4.2. Optimised Topics in the Selected Topic Model

Table 4 shows the inferred, optimised topics and their associated word clusters for the selected topic model (i.e. Model 2). Of the 20 word clusters generated from the model, 14 unique topics were inferred; two word clusters were interpreted as referring to the same topic, and four word clusters were indicated as "Not Applicable" (N.A.) as no meaningful topic could be inferred. After optimising the inferred topics by reviewing the selected posts, 13 unique topics were found. They are identified as a mixture of clinical and non-clinical topics. The clinical topics (CT) include "palliative care", "rheumatology", "evidence-based medicine", "statins use", "vitamin use", "vaccines", "women's health check", and "fibromyalgia"; the non-clinical topics (NCT) include "patient fees", "training", "prescriptions", "policy", and "workload".

Table 4. The optimised topics for the selected model

Cluster	**Topic weight**	**Topic words**	**Inferred topic**	**Optimised topic**
1	0.19594	pay medicare bulk work charge fee private service money government	Bulk billing	Patient fees (NCT)
2	0.18415	prescription script pharmacy addict drug pharmacist pbs authority day write	Prescriptions	Prescriptions (NCT)
3	0.0322	food car house fridge eat poor store hot change balance	N.A.	N.A.
4	0.1898	patient pay medicine stress society reduce current government rate finance	Health cost	Patient fees (NCT)
5	0.08499	care nurse hour visit service palliative home medical provide rural	Palliative care	Palliative care (CT)
6	0.04848	medical profession racgp public doctor ahpra nurse health wrote report	Training	Training (NCT)
7	0.05038	point vaccine medicine understand body view generate form base suggestion	Vaccines	Vaccines (CT)
8	0.03467	refer comment expert issue interest person lack call present programme	N.A.	N.A.
9	0.07868	restrict country year work moratorium hospital system area dws law	Policy	Policy (NCT)
10	0.10449	trial statin evidence effect prevent group side benefit interest study	Statins use	Statins use (CT)
11	0.07429	effect level side dose disease vitamin high symptom difference drug	Vitamin use	Vitamin use (CT)

12	0.09575	pain inject joint muscle guidance bursa knee stretch tear elbow	Elbow bursa treatment	Rheumatology (CT)
13	0.07989	evidence point medicine base comment treatment understand body view ahpra	Evidence-based medicine	Evidence-based medicine (CT)
14	0.07777	examination check breast pap women history year present cancer diagnose	Women's health checks	Women's health checks (CT)
15	0.1173	human organ end therapy cell central protect age create state	N.A.	N.A.
16	0.0878	risk diabetes cholesterol disease level statin calculate exercise hdl year	Statins use	Statins use (CT)
17	0.03857	pain chronic toe relief chest attached post bit metal fibromyalgia	Fibromyalgia	Fibromyalgia (CT)
18	0.17284	patient practice doctor person time medical work care health hour	Work hours	Workload (NCT)
19	0.06752	practice hospital general training specialist nurse year medical head base	Training	Training (NCT)
20	0.02064	record request summary advice provide inform letter email initial legal	N.A.	N.A.

As shown in Table 4, there are more clinical than non-clinical topics identified from the dataset. However, the weights of the topics imply that non-clinical topics were more frequently discussed than clinical ones.

With regards to the clinical topics, palliative care, rheumatology, and evidence-based medicine appeared to generate some in-depth discussion among the participants. By inspecting a number of specific posts on the topics relating to women's health checks, fibromyalgia, the use of statins, vaccines, and vitamin, we noted that the participants were interested in benchmarking their practices. This is understandable as clinical practice can be conducted differently in different places; OSNs have been found to enable health professionals to share different ways of performing the same practice and benchmark the most effective one [16].

The non-clinical topics identified from this dataset are mostly controversial (include policy, workload and patient fees). This finding is consistent with previous studies that have demonstrated health professionals are particularly interested in discussing controversial topics in an OSN [17]. In addition, the participants were keen to keep themselves up-to-date on advanced information and news in the field; this is reflected in the topics relating to policies, training, and prescriptions.

4.3. Topic Evaluation

The 13 identified topics were evaluated using F-measure against 40 randomly selected posts from the dataset. The Accuracy, Precision, Recall, and F-score of the topic model were 0.53, 0.63, 0.70, and 0.66 respectively. The Accuracy of 0.53 indicates that the topic model is likely to capture 53% of the topics in any randomly selected posts. The F-score of 0.66 informs that the topic model correctly captures 66% of the overall topics in this random selection of 40 posts.

5. Conclusions and Future Work

OSNs have been increasingly used by health professionals to share medical knowledge and experience. However, there is a lack of understanding about how health professionals learn in OSNs, making it difficult to design and facilitate this type of learning. This study contributes towards understanding their learning resource and development needs in OSNs by demonstrating the use of topic modelling to identify the topics of interest that emerge from an online discussion forum of health professionals.

The evaluation of the topic model was performed using F-measure. The F-score of 0.66 informs that the topic model is not optimal but correctly captures 66% of the overall topics in a random selection of 40 posts. This suggests that topic modelling could be used to identify the emerging learning topics from the large amount of textual dialogue generated in an OSN. As we have found no previous work on topics discussed by an OSN for health professionals to compare our results with, it is inconclusive whether the topics we identified are typical or atypical of those discussed by health professionals. However, the results suggest that the health professionals in this OSN

are interested in knowing or discussing clinical topics relating to palliative care, rheumatology, evidence-based medicine, women's health checks, fibromyalgia, the use of statins, vaccines, and vitamins, as well as non-clinical topics relating to prescriptions, patient fees, policy, workload, and training.

Identifying topics using this method could provide education designers and OSN operators with guidance on facilitating online discussion that is most relevant to the learning needs of health professionals. In this OSN, it has been found that non-clinical topics were more frequently discussed than clinical ones by the health professionals. Without knowing the context, we could not support having non-clinical topics as the main focus of their online discussion, but it is important to consider how to help health professionals to deal with the challenge of keeping themselves up-to-date on non-clinical and work-related information. In addition, it might be worth considering proposing common clinical topics relating to their clinical practices that allow them to share practical and experiential knowledge and meet the needs for benchmarking.

A limitation of this study is that considering the overall activity in the discussion forum within this OSN, data were analysed very selectively. Due to limitations of the data source, passive users (i.e. those who learn by reading but do not participate in any discussion) were not tracked in our study, which means the topics identified only apply to the active participants of this OSN.

In a future study, we plan to include additional meta-data to fit into the topic model, for example, including the identity of the authors enables us to investigate author similarity based on their discussion of topics. This will help to group health professionals who may have similar learning needs. Furthermore, understanding of the learning context (e.g. goals, tasks, preference, interests, and constraints) enhances the interpretation of the identified topics.

6. References

1. Cheston, C.C., T.E. Flickinger, and M.S. Chisolm, *Social media use in medical education: a systematic review.* Academic Medicine, 2013. **88**(6): p. 893-901.
2. Sandars, J., P. Kokotailo, and G. Singh, *The importance of social and collaborative learning for online continuing medical education (OCME): directions for future development and research.* Med Teach, 2012. **34**(8): p. 649-652.
3. Institute of Medicine, *Redesigning Continuing Education in the Health Professions*. 2010, National Academies Press: Washington, DC.
4. Stewart, S.A. and S.S.R. Abidi, *Using Social Network Analysis to Study the Knowledge Sharing Patterns of Health Professionals Using Web 2.0 Tools.* Biomedical Engineering Systems and Technologies, 2013. **273**: p. 335-352.
5. Li, X., et al., *Analysing Health Professionals' Learning Interactions in Online Social Networks: A Social Network Analysis Approach*, in *Health Informatics New Zealand Conference*. 2015: Christchurch, New Zealand.
6. De Laat, M., *Network and content analysis in an online community discourse*, in *Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community*. 2002, International Society of the Learning Sciences: Boulder, Colorado. p. 625-626.
7. Schrire, S., *Knowledge building in asynchronous discussion groups: Going beyond quantitative analysis.* Computers & Education, 2006. **46**(1): p. 49-70.
8. Blei, D.M., *Probabilistic topic models.* Communications of the ACM, 2012. **55**(4): p. 77-84.
9. Tobarra, L., et al. *Discovery of interest topics in web-based educational communities*. in *Computers in Education (SIIE)*. 2012. Andorra la Vella: IEEE.
10. Portier, K., et al., *Understanding topics and sentiment in an online cancer survivor community.* JNCI Monographs, 2013. **47**: p. 195-198.
11. Ezen-Can, A., et al. *Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach*. in *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. 2015. ACM.
12. McCallum, A.K., *MALLET: A Machine Learning for Language Toolkit*. 2002.
13. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent Dirichlet Allocation.* The Journal of Machine Learning Research, 2003. **3**: p. 993-1022.
14. Chang, J., et al. *Reading tea leaves: How humans interpret topic models*. in *Advances in neural information processing systems*. 2009.
15. Van Rijsbergen, C.J., *Information retrieval*. 2nd ed. 1979, Butterworth.
16. Millar, B., K. Ho, and A.-M. Carvalho, *Social media to support physician practice and CPD: Opportunities, issues, and an emergency medicine case study.* BCMJ, 2016. **58**(2): p. 94-96.
17. Panahi, S., J. Watson, and H. Partridge, *Social media and physicians: exploring the benefits and challenges.* Health Informatics Journal, 2014: p. 1460458214540907.