An Emotion Oriented Topic Modeling Approach to Discover What Students are Concerned about in Course Forums

Zhi Liu[a,b], Tai Wang[a], Niels Pinkwart[c], Sannyuya Liu[a,b], Lingyun Kang[a]
[a]National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, China
[b]National Engineering Laboratory for Educational Big Data, Central China Normal University, Wuhan, China
[c]Department of Computer Science, Humboldt University of Berlin, Berlin, Germany
{zhiliu, wangtai}@mail.ccnu.edu.cn, niels.pinkwart@hu-berlin.de, lsy5918@mail.ccnu.edu.cn,
kanglingyun@mail.ccnu.edu.cn

***Abstract*—Course forums offer an interactive channel for learners to express opinions and feedback, which contain valuable emotions and topic information towards courses. In this paper, we propose an emotion oriented topic probabilistic model that can be used to calculate distributions of emotion-topic over words to discover what students are most concerned about. An experiment on real-life data indicates that students had a positive attitude for knowledge applications, a negative experience for the learning system, and expressed confusion about the final exam. We also visualize the temporal trends of emotions of the whole group and the groups with different levels of achievement. The proposed model has a potential in discovering students' emotions in their feedback, thus improving the online learning experience, and identifying at-risk students timely.**

Keywords-learning analytics; course forum; topic probabilistic model; emotion prior knowledge

I. Introduction

The emergence of Massive Open Online Courses (MOOCs) and Small Private Online Courses (SPOCs) has prompted the rise of "big data" in education. A lot of textual data is being generated in online course forums. This forum data is a documentation of students' interactive learning processes, which can allows for in-depth observations and analyses of students' authentic voices in a non-invasive way [1][2]. Like consumers, students may deliver various feedback to evaluate everything they experience online. They share learning feelings about online courses, which provides many opportunities to discover students' emotional states. However, the unstructured textual data may pose a difficulty for teachers who want to understand the feedback. Therefore, it is helpful to summarize latent opinions to uncover what students are concerned about in forums.

In recent years, topic and emotion analyses in texts have gain much interest of researchers in the field of learning analytics due to the potential function for students' opinions in predicting their drop-outs. Ramesh et al. developed the weakly supervised aspect-sentiment joint model to extract topic words and to detect sentiment polarities of forum posts, which helps instructors select a subset of critical posts to address main reported issues by students [3]. Wen et al. calculated collective sentiments under the four predefined topics (course, lecture, assignment and peer-assessment) from different forums to track students' trending opinions about courses and course tools [4]. Brinton proposed a unified topical model for the discussion threads, extracting keywords from forum posts to identify course content-related discussions [5]. Liu et al. proposed a deterministic emotional information based topic model to explore advantages and disadvantages in a course and calculate the overall popularity of a course [6]. To sum up, the joint modeling of topic and emotion has become a striking trend in the learning analytics community. However, these modeling approaches require either lots of predefined rules for associating topics with emotions, or post-processing to connect topics with emotions, which is typically labor-consuming.

II. Course Discussion Dataset

The dataset used in this study comes from the two SPOC discussion forums of "Introduction to Psychology", a degree course open at a normal university of Central China. The dataset comprises a total of 10,870 posts by 752 undergraduate students across the first and second semesters of academic year 2014-2015. The participants include 719 registered and 33 non-registered students. The average score of the registered students is 75.97 in final exams (on a scale of 0-100) with a variance of 13.87.

III. Emotion-Oriented Topic Modeling

A. Modeling Association between Emotion and Topic

In the emotion-oriented topic model (EoTM), emotion is viewed as a heuristic factor to generate topics. Faced with a large-scope discussion, a teacher may prefer to know the primary objects which students are interested/disinterested in, happy/unhappy with and confused with rather than what emotion they express for each aspect in online learning. We assume that, before writing a post, a student has an intuitive feeling corresponding to the distribution of emotions for his/her online learning experience, e.g., 50% satisfied, 30% unsatisfied and 20% confused. And his/her evaluative ideas could be modeled by a probabilistic distribution of the topics for each emotion, e.g., 40% for the assignment difficulty, 30% for the lecture styles, 30% for the schedule under the negative emotion. Then the student decides, for each sentence, an emotion to express and an associated topic. Thus, the modeling process of EoTM is shown as follows:

2161-377X/18/$31.00 ©2018 IEEE
DOI 10.1109/ICALT.2018.00047

```
1. Draw a word distribution $\phi_{ez} \sim Dirichlet(\beta_e)$ for each pair of
   emotion $\mathbf{e}$ and topic $\mathbf{z}$
2. For each post $d$,
   (a) Draw the post's emotion distribution $\pi_d \sim Dirichlet(\gamma)$
   (b) Draw a topic distribution $\theta_{ds} \sim Dirichlet(\alpha)$ over each emotion
       $e \in \{pos, neg, conf\}$
   (c) For each sentence
       i.Sample an emotion label $j \sim Multinomial(\pi_d)$
       ii.Given emotion $j$, sample an topic label $k \sim Multinomial(\theta_{dj})$
       iii.Generate words $w \sim Multinomial(\phi_{jk})$
```

Figure 1. Generative process of EoTM.

Here *α*, *β*, *γ* denotes the Dirichlet prior for document-emotion-topic distribution *θ*, emotion-topic-word distribution *ϕ* and document-emotion distribution *π*, respectively.

B. *Parameter Inference*

We adopt Gibbs sampling to infer the hidden parameters such as *θ*, *ϕ*, *π* in EoTM. According to the Bayes conditional independence criterion, the joint probability of topics, emotion and words can be formed as follows:

$$P(\mathbf{w}, \mathbf{e}, \mathbf{z} | \alpha, \beta, \gamma) = P(\mathbf{w} \mid \mathbf{e}, \mathbf{z}, \beta) \cdot P(\mathbf{e} \mid \gamma) \cdot P(\mathbf{z} \mid \mathbf{e}, \alpha) \quad (1)$$

At each transition stage of the Markov chain, given all known conditional variables, the posterior distribution is estimated by iteratively sampling the variables **w** (word), **e** (emotion), and **z** (topic). The emotion and topic of the *i*-th sentence are calculated using the conditional probability:

$$P(e_i = j, z_i = k \mid \mathbf{e}_{\mathbf{-i}}, \mathbf{z}_{\mathbf{-i}}, \mathbf{w})$$
$$\propto \frac{C_{dj}^{DE} + \gamma_j}{\sum_{j'=1}^{E} C_{dj'}^{DE} + \gamma_{j'}} \cdot \frac{C_{djk}^{DET} + \alpha_{jk}}{\sum_{k'=1}^{T} C_{djk'}^{DET} + \alpha_{jk'}} \cdot \frac{C_{djk}^{STW} + \beta_{jk}}{\sum_{w'=1}^{V} C_{jkw'}^{STW} + \beta_{jw'}} \quad (2)$$

In this formula, the right three items correspond to the three distributions π_{dj}, θ_{djk}, ϕ_{jkw} of the *i*-th sentence in the *d*-th post, among which ϕ_{jkw} is utilized to indicate what students are most concerned about in discussion forums.

C. *Incorporating Emotion Prior Knowledge*

To precisely identify the emotion of each sentence, we adopt the emotion lexicon constructed in [7] to match terms indicating positive and negative moods or appraisals. Besides, we build a word set indicating confused emotion. The emotion lexicon can be illustrated as follows:

TABLE I. DISTRIBUTION OF EMOTION LEXICONS IN FORUM DATA

Category	Words	Matched terms	High-frequency terms (top 5)
Positivity	9586	47782	“正式/formal”, “深远/profound”, “愉快/happy”, “很充沛/very abundant”, “聚精会神/concentrate”
Negativity	12871	30924	“压制/suppress”, “非议/criticism”, “闭塞/closed”, “不深刻/not profound”, “乏味/tedious”
Confusion	934	6500	“可不可以/can or not”, “莫名其妙/mysterious”, “迷茫/perplexed”, “似是而非/paradoxical”, “迷失/lost”

In the initialization stage, the emotion lexicon, as a prior knowledge, is used to label the emotion of each sentence. Thus, if one word exists in the lexicon, it would be assigned with the corresponding emotion label. The emotion label of one sentence is determined by the majority voting on occurrence frequencies of different emotion categories of words. If there is an equal number of words for each emotion or there are not any emotional words in a sentence, the emotion label is sampled by Equation (2).

IV. EXPERIMENTAL RESULTS

In our experiment, we set the number of topics to 60 and set *α*, *β*, *γ* as the symmetric parameters, i.e., 0.1, 0.01 and 1. The number of iterations for Gibbs sampling is 500.

A. *What Students Were Most Concerned About*

For each emotion, the topic with the highest probability is described with representative words in Table II.

Positive-T1 means a topic about how to keep a good mentality, students tend to positively use psychological knowledge to adjust personal mentality in real life.

Negative-T25 implies some technical problems such as videos' playing, system errors, etc. in Cloud Classroom, a SPOC learning system deployed in the university.

Confusion-T38 indicates a confused emotion about the final exam, students might be concerned about the final exam format and potential questions.

TABLE II. PARTIAL PROBABILISTIC DISTRIBUTION OF EMOTI-TOPICS WITH HIGHEST PROBABILITIES

Emoti-topic label	Top 10 words with highest probabilities
Positive-T1	好/good (0.037), loveliness (0.023), 生活/life (0.023), 乐观/optimism (0.019), 积极/positive (0.017), 心态/mentality (0.016), 调整/adjust (0.015), 培养/foster (0.011), 赞同/agree (0.011), 方法/method (0.011)
Negative-T25	没时间/no time (0.058), 老师/teacher (0.034), 作业/assignment (0.030), 章节/chapter (0.021), 视频/video (0.019), 重新/repeatedly (0.019), 云课堂/Cloud Classroom (0.016), 问题/problematic (0.015), 系统/system (0.011), 错误/error (0.010)
Confusion-T38	考试/exam (0.094), 考/test (0.070), 期末/end of semester (0.035), 老师/teacher (0.026), 一样/same (0.026), 习题/exercise (0.026), 第九章/the 9th chapter (0.026), 闭卷/closed book exam (0.026), 不知道/do not know (0.021), 很难/very difficult (0.016)

Note: the emotional terms are marked in red.

B. *Temporal Trends of Emoti-Topics*

Due to space limitation, we only select the forum data of the second semester to visualize dynamics of the above three pairs of *emoti-topics*. Thus, the document-emotion-topic probabilities in the same teaching week are averaged to form an overall temporal emotion-topic distribution as follows:

Figure 2 indicates that the negative emotion related with system operations mainly occurs at the initial stage of the semester, and after that the negative emotion has been at a very low level. The positive emotion related with mental adjustment seems relatively stable during a semester. The confused emotion about the final exam shows a plausible increase from the 12th week to the end of the semester.

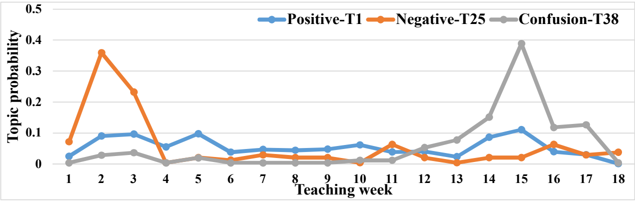

Figure 2. Temporal trends of three emoti-topics among students.

C. *Temporal Emoti-Topics of Different Achieving Students*

We divided all forum participants into four groups (high-, medium-, low-achieving and non-registered groups) according to their achievement levels in the final exam, which is defined by the High-Low Discrimination index [8]. Among them, the non-registered group only engaged in forums and did not take exams. Figure 3-5 show the temporal trends of the four groups in the three pairs of *emoti-topics* with the highest probability in the corresponding emotion-topic distribution. For the probabilities of Positive-T1 and Negative-T25, the high-achieving group displays the higher level, and the non-registered group is in the lowest level. However, the low-achieving group shows the highest level of confusion at the end of the semester, suggesting that this group has had a difficulty in preparing for exams. This is similar to the findings in MOOC forums [9]. This phenomenon is worthy of teachers' attention, as a timely response to the needs of potential at-risk students may contribute to improve their learning performance.

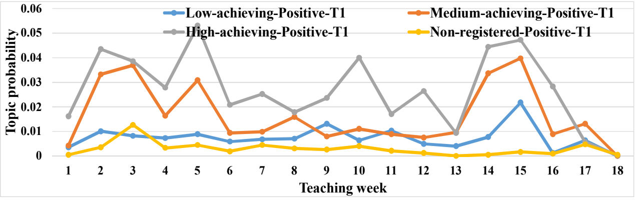

Figure 3. Temporal trends of Positive-T1 for the four groups.

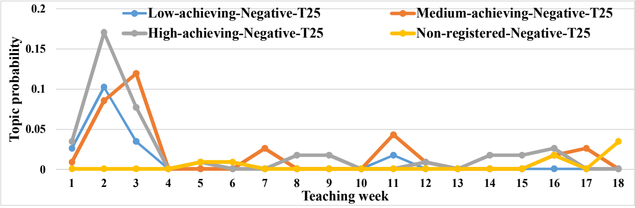

Figure 4. Temporal trends of Negative-T25 for the four groups.

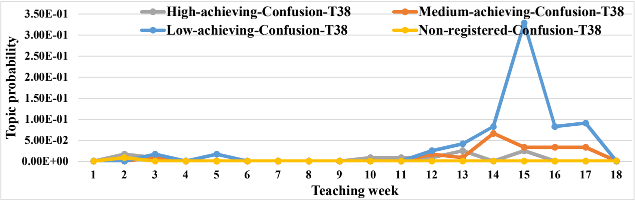

Figure 5. Temporal trends of Confusion-T38 for the four groups.

V. Conclusion

In this study, we proposed a joint probabilistic model, called EoTM that incorporates emotion prior knowledge to calculate the emotion-specific topic distribution over forum posts, to discover what students are most concerned about. The critical topics associated with positive, negative and confused emotions can be sufficiently detected to indicate students' interests in psychological knowledge application, demands for improving the user experience of the system, and difficulties in preparing for the final exam. Especially for negative and confused emotions, EoTM has a potential value to help teachers identify at-risk students in advance by a dynamic tracking of *emoti-topics* during a semester.

Acknowledgment

This work is supported by the Research Funds from National Natural Science Foundation of China (Grant No. 61702207, L1724007, 31600918), China Scholarship Council (Grant No. 201706775022), National Key Research and Development Program of China (Grant No. 2017YFB1401303), Project of Humanities and Social Sciences (Grant No. 16YJC880052), National Social Science Fund Project of China (Grant No. 14BGL131) and BMBF (German Ministry of Education and Research) project LISA (Grant No. 16SV7535).

References

[1] H. B. Shapiro et al., "Understanding the massive open online course (MOOC) student experience: An examination of attitudes, motivations, and barriers," Comput Educ., vol. 110, pp. 35-50, 2017.

[2] P. Rodriguez, A. Ortigosa, and R. M. Carro. "Extracting Emotions from Texts in E-learning Environments," Proceedings of the Sixth International Conference on Complex, Intelligent, and Software Intensive Systems, 2012, pp. 887-892.

[3] A. Ramesh, D. Goldwasser, B. Huang, H. Daumé, and L. Getoor. "Modeling learner engagement in MOOCs using probabilistic soft logic," NIPS Workshop on Data Driven Education. 2013, pp. 21-62.

[4] M. Wen, D. Yang, C. P. Rosé. "Sentiment analysis in MOOC discussion forums: What does it tell us," Proc. of 7th Intl. Conf. on Educational Data Mining, 2014, pp. 130-137.

[5] C. G. Brinton et al., "Learning about social learning in MOOCs: From statistical analysis to generative model," IEEE Trans. Learning Tech., vol. 7, no. 4, pp. 346-359, 2014.

[6] Z. Liu et al., "Emotion and associated topic detection for course comments in a MOOC platform," Proc. of 5th Intl. Conf. on Educational Innovation through Technology (EITT), 2016, pp. 15-19.

[7] Z. K. Yang, Z. Liu, S. Y. Liu, M. Lei, and W. T. Meng. "Adaptive multi-view selection for semi-supervised emotion recognition of posts in online student community," Neurocomputing, vol. 144, pp. 138-150, 2014.

[8] T. L. Kelley, "The selection of upper and lower groups for the validation of test items," J. Educ. Psychol., vol. 30, no. 1, pp. 17-24, 1939.

[9] D. Yang, R. Kraut, and C. P. Rosé, "Exploring the Effect of Student Confusion in Massive Open Online Courses," Journal of Educational Data Mining, vol. 8, no. 1, pp. 52-83, 2016.