Topic Sentiment Analysis in Online Learning Community from College Students

Kai Wang, Yu Zhang[†]

Bengbu Medical College, Bengbu, Anhui, China

Abstract

Purpose: Opinion mining and sentiment analysis in Online Learning Community can truly reflect the students' learning situation, which provides the necessary theoretical basis for following revision of teaching plans. To improve the accuracy of topic-sentiment analysis, a novel model for topic sentiment analysis is proposed that outperforms other state-of-art models.

Methodology/approach: We aim at highlighting the identification and visualization of topic sentiment based on learning topic mining and sentiment clustering at various granularity-levels. The proposed method comprised data preprocessing, topic detection, sentiment analysis, and visualization.

Findings: The proposed model can effectively perceive students' sentiment tendencies on different topics, which provides powerful practical reference for improving the quality of information services in teaching practice.

Research limitations: The model obtains the topic-terminology hybrid matrix and the document-topic hybrid matrix by selecting the real user's comment information on the basis of LDA topic detection approach, without considering the intensity of students' sentiments and their evolutionary trends.

Practical implications: The implication and association rules to visualize the negative sentiment in comments or reviews enable teachers and administrators to access a certain plaint, which can be utilized as a reference for enhancing the accuracy of learning content recommendation, and evaluating the quality of their services.

Originality/value: The topic-sentiment analysis model can clarify the hierarchical dependencies between different topics, which lay the foundation for improving the accuracy of teaching content recommendation and optimizing the knowledge coherence of related courses.

Keywords Online learning community; Topic detection; Sentiment analysis

Citation: Wang, Kai, and Yu Zhang. "Topic sentiment analysis in online learning community from college students." *Journal of Data and Information Science*, vol. 5, no. 2, 2020, pp. 33–61. <https://doi.org/10.2478/jdis-2020-0009>

Received: Sep. 19, 2019

Revised: Jan. 8, 2020

Accepted: Mar. 23, 2020

[†] Corresponding author: Yu Zhang (E-mail: 475870652@qq.com).

1 Introduction

Due to the rapid development of Web 2.0 and social media, the Online Learning Community (OLC) is increasingly being utilized by academic institutions to create a more convenient learning environment (Fariza, 2019; Wu, Hsieh, & Yang, 2017). OLC establishes a virtual social form through teaching, research and other activities, with interactive learning, collaborative learning and independent learning. In summary, OLC consists of three basic elements: technology, teaching, and academic sentiment interaction. The purpose of improving academic sentiment interaction is to cultivate learners' sense of belonging to the community, so that learners are willing to stay in the community for a long time and maintain learning motivation at a high level (Cho, Kim, & Choi, 2017). Academic sentiments are generally hidden in the text records of learning community activities, such as documents, statements, and sentences. Through the techniques of sentiment analysis, weight calculation, and semantic understanding, the sentiment experience related to learning processes can be observed. In this regard, academic sentiment mining (AEM) can gain insights from the comments in OLC to analyze the factors affecting learning outcomes, which is of great significance for the improvement of teaching theories and methods (Kohoulat et al., 2017).

In the realization of AEM, much attention has been paid towards discoveries of the expressed sentiments in the applications of academic recommender system (Kaklauskas, Zavadskas, & Seniut, 2013), opinion leader identification (Li, Ma, & Zhang, 2013), forum topic mining (Cheng et al., 2015; Colace, Santo, & Luca Greco, 2014), and community quality analysis (Ghiasifard et al., 2015), etc. However, all the available methods mentioned above are unable to highlight the identification and visualization of topic sentiment based on learning topic mining and sentiment clustering at various granularity-levels. In the light of recent studies, current paper aims at constructing topic sentiment analysis in OLC. Our ultimate goal is to obtain a list of terms relevant to some learning topic and to visualize the association relationships based on the sentiment classification in an interactive way.

To access a certain plaint set without checking all the document information, current paper aims at highlighting the identification and visualization of topic sentiment based on learning topic mining and sentiment clustering at various granularity-levels. To enact this need, we first proposed a topic analytics method with actual data sets from the website (www.icourses.cn), to construct the documents-topics probability matrix and the topics-terms probability matrix. Secondly, the topic sentiment distances are measured on the basis of topic-clustered concept lattice. Then, the sentiment polarity can be calculated with the help of the sentiment

dictionary and domain context. Finally, a set of formal concepts containing topic-terms as well as a set of association rules are generated and visualized.

The following proposals are given in the paper.

1. A topic analytics method is proposed to analyze and extract potential topics in OLC.
2. On the basis of college students' feedback, a novel approach is developed to identify the topic sentiment by measuring the sentiment distance.
3. In addition, the hierarchical and associated relationships as well as the granularity of sentiment information are obtained.

The rest of the paper is constituted as follows: Section 2 introduces a related survey of methodologies in the area of topic detection, sentiment analysis, and sentiment concept clustering. Section 3 gives introductions about Latent Dirichlet Allocation (LDA) and Formal Concept Analysis (FCA). The topic sentiment analysis method is proposed in Section 4. Section 5 contains the illustration and implementation of the proposed method followed by the results and discussions given in Section 6. Section 7 presents the conclusions and future work followed by the acknowledgment and references.

2 Related work

The OLC contains a large amount of information, which can be divided into two parts: learning resources and student review information (Shea, Li, & Pickett, 2006). How to accurately extract the topic of students' attention and make corresponding optimization according to the sentiment of students' comments has become a key link to improve the quality of community service and enhance the learning effect of students. To achieve this goal, many scholars have carried out extensive and in-depth research, which can be summarized into three main steps: topic detection (Lu et al., 2013), sentiment analysis (Nan & Wu, 2010), and sentiment concept clustering (Pappas et al., 2017). In this regard, various investigations and studies have been put forward to optimize the quality of data mining in OLC. In order to explain the proposed method in an understandable way, this section focuses on the detection methods of community topics and analyzes different sentiment distribution models according to the corresponding topics.

2.1 Topic detection for online community

The topic detection model is a kind of probability generation model for text content by simulating the human mind process to find the best topic set and its vocabulary. The existing topic models mainly include: latent semantic analysis

(LSA) (Martínez, 2015), probabilistic semantic based indexing model (PLSI) (Parvathy, 2016) and Latent dirichlet allocation (LDA) (Yue, Barnes, & Jia, 2017).

The LSA method implements the representation of documents on low-dimensional implicit semantic spaces by introducing semantic dimensions. Although the model is capable of constructing text representation without dictionary spaces, the basis of the LSA methodology is still derived from linear algebra, generating huge amount of negative numbers in various dimensions. To deal with this problem, Hofmann proposed a probabilistic semantic based indexing model (PLSI) (Hofmann, 1999). The aim of PLSI is to emphasize the semantic interpretability of topic text based on implicit semantic indexing, which is unable to deal with the over-fitting problems caused by massive text. In order to improve the parameters from PLSI, which cannot be linearly changed as the document set grows, Blei proposed the LDA (Latent Dirichlet Allocation) model to retrieve potential topics, representing high-dimensional word space with low-dimensional topic space (Blei, Ng, & Jordan, 2003). The LDA model is a multi-layer unsupervised Bayesian network that has been widely used to mine document subject knowledge. The LDA-based approach for online community can be summarized into two categories. The first aspect is to identify similar topics under different time segments and analyze the evolution trends. Chu and Li (2010) proposed a method to realize the evolution of the topics. They utilized the original corpus for topic classification. Nagori developed a content-based recommended system to personalize the e-learning systems (Nagori & Aghila, 2012). They exploited the topic model by introducing the similarity metrics. Yang (Yang, Zhang, & Shi, 2014) adjusted the priori parameters of the model to find changeable topics in the text. Ge extracted the hidden micro blog topics to emerge topics that need to be expressed in the community (Ge, Chen, & Du, 2013). The second aspect is the combination with other models to enhance deep semantic relationships of the topic. Santosh, Vardhan, and Ramesh (2016) focused on the analysis of the feature attributes of online product reviews. They proposed the LDA model to obtain the feature keywords of the product and combined the feature ontology tree (FOT) to improve the accuracy of subject detection. Cerulo and Distant (2013) obtained a topic-terms matrix by developing a topic recognition model, which was utilized to form a formal context to constructing a theme concept lattice for topic-driven navigation. Zhong et al. (2018) designed an evaluation framework for the quality of student comments in online communities. They considered the dimensional characteristics of online commentary data quality, and constructed a set of topic features. To sum up, the current methods focus on the evolution analysis of community topic mining, semantic relationship enhancement, and probabilistic topic modeling, ignoring the hierarchical relationship between topics and sentimental analysis of students' feedback.

2.2 Sentiment analysis for online students

In the process of topic detection, adding sentiment analysis can identify the sentiment changes from the online students implied in the topic. Therefore, it is necessary to identify sentiment distributions according to the corresponding topic. Sentiment analysis, also known as opinion mining, is the process of analyzing, processing and classifying subjective texts with sentiment techniques. At present, the mainstream sentiment analysis methods can be divided into three categories. The first aspect is to analyze the text by constructing a sentiment dictionary, which mainly relies on the qualities of sentiment lexicons with specific semantic rules. Pointwise Mutual Information (PMI) and Latent Dirichlet Allocation (LDA) are often used in constructing sentiment lexicons, among which PMI can be used to judge the sentiment tendency of words, while LDA is utilized to extract sentiment words from corpus (Li, Ba, & Huang, 2015). Turney and Littman (2003) developed the PMI algorithm to extend the sentiment dictionary, and then the semantic polarity algorithm is proposed to analyze the sentiment tendency of the text, which improves the accuracy of text data classification. Yang, Peng, and Chen (2014) proposed a LDA-based method to constructing a specific domain sentiment dictionary on the basis of the existing public sentiment dictionary, where the extracted topic words are viewed as a priori knowledge from the corpus. The second aspect is focused on mining the sentiment features of the text based on Machine Learning (ML), such as Support Vector Machine (SVM) (Liu, Bi, & Fan, 2017), Naive Bayes (NB) (Shirakawa et al., 2017), Maximum Entropy (ME) (Ficamos, Yan, & Chen, 2013). Vinodhini (2014) designed a hybrid formwork of SVM and principal component analysis (PCA) to improve the sentiment classification accuracy by reducing the complexity of the sentiment mining model. Mertiya and Singh (2016) proposed an unsupervised polarity selection method to determine the polarity of tweets via merging NB and adjective analysis theory. Xie et al. (2017) extracted the seed sentiment words from Wikipedia by using probabilistic latent semantic analysis, which are used as the input matrix of the ME model. Meanwhile, to classify sentiment, they used entropy classification theory to select sentiment features. In addition, the last aspect is the deep-learning based approach by converting word embedding into a text vector to extract deep sentiment features, which mainly includes Convolutional Neural Networks (CNN) and Recurrent Neural Network (RNN). Shin, Lee, and Choi (2017) integrated lexicon embeddings, attention mechanisms into CNN to analyze sentiment features with less noisy words. Ethemet, Aysu and Fazli (2018) proposed a cross-language emotion analysis model, which can realize sentiment analysis based on CNN under the condition of small corpus. Although many researchers have put a lot of efforts into improving the sentiment

Research Paper

classification of online communities for practical work, there is still a lack of evaluation in the sentiment unit combination, especially when it comes to the OLC. Since the sentiment analysis of students' learning is closely related to the context where the topic is located, it is necessary to establish a set of association rules with contextual awareness. To enact this need, we introduce formal concept analysis theory into online sentiment analysis by exploring the sentiment association rules between students.

To sum up, current paper tries to make improvements in two ways. Firstly, the granularity analysis of learning topic for visualizing the hierarchical relationships is considered. Afterwards, we stay focus on finding the negative sentiment form students' comments on the basis of sentiment scoring calculation to form the basic association rule sets.

3 Introduction to LDA and FCA

3.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a three-layer Bayesian probability network, which assumes that documents in the corpus select a topic based on a certain probability, and each topic also selects a term based on a certain probability. Therefore, a document is a mixture of multiple topics, and a topic is also a mixture of multiple terms. Suppose the topic distribution vector in the document is θ , the topic term distribution is φ , z is the topic, and w is the term, where θ and z are the implicit variables, w is the explicit variable. The topic recognition process of the LDA model is shown in Figure 1 (Blei D. M., Ng A. Y., & Jordan M. I., 2003). Topics-Terms Matrix and Topics-Documents Matrix can be mined in the end.

Figure 1. Schematic diagram of topic generation based on LDA model.

3.2 Formal concept analysis

Formal concept analysis (FCA) is a hierarchical concept construction theory based on Galois connection, which is utilized to describe the domain knowledge in depth on the basis of the mapping relationships between objects and attributes. The

FCA theory consists of four basic notions of formal context, formal concept, partial ordering, and concept lattice. To further analyze the collections of documents in OLC by referring to Ren's paper (Ren, Ling, & Yao, 2018), four definitions are given separately.

Definition 1 (Formal context) (Wei et al., 2019) A formal context is represented as a triple $K=(S, T, I)$, where I is a binary relation represented as ${}_SI_T$ between a set of objects $S=\{s_1, \dots, s_i, \dots, s_N\}$ (a collection of documents) and a set of attributes $T=\{t_1, \dots, t_j, \dots, t_K\}$ (a collection of topics).

Definition 2 (Formal concept) (Wei et al., 2019) Let us suppose $K=(S, T, I)$ be a formal context, $X \subseteq S$, $Y \subseteq T$, and (X, Y) be a formal concept if it satisfies the following conditions:

$$\forall x \in X, \forall y \in Y \Rightarrow (x, y) \in I \quad (1)$$

$$\forall x \notin X \Rightarrow \exists y \in Y, (x, y) \notin I \quad (2)$$

$$\forall y \notin Y \Rightarrow \exists x \in X, (x, y) \notin I \quad (3)$$

The inheritance relationship between different formal concepts can be utilized to construct a complete concept lattice through partial order relations, which is defined as follows.

Definition 3 (Partial ordering) (Zhang, Wei, & Qi, 2005) Let us suppose (X_1, Y_1) and (X_2, Y_2) be two formal concepts. The partial ordering relation between the two formal concepts is valid on condition that $X_1 \subseteq X_2$ or $Y_1 \subseteq Y_2$. In that case, (X_1, Y_1) is called the subconcept of (X_2, Y_2) .

Definition 4 (Concept lattice) (Zhang, Wei, & Qi, 2005) Let us suppose $\prec$ be the set of partial orderings among the whole formal concepts. $H(S, T, \prec)$ is defined as a concept lattice based on the formal context $K=(S, T, I)$.

Current paper focuses on the following steps to demonstrate the advantages of utilizing LDA for opinion mining: Firstly, when traditional machine learning methods are applied to sentiment classification, the classification effect is unstable, and most of them are supervised methods, which require a certain number of labeled training samples. The manual labeling process is relatively time-consuming and labor-intensive with poor field portability. Therefore, unsupervised learning algorithms have become an important research direction for sentiment analysis of online reviews. However, although the existing thematic sentiment hybrid model can extract both the subject and sentiment information of the document at the same time, the effect of the model's sentiment classification and the stability are not ideal due to the local negation and the number of subjects in the subjective document. In fact, text sentiment classification is still essentially a text classification problem.

The innovations of this paper can be summarized as the following two points: 1) Modeling the specific domain knowledge of college students' online learning communities, as well as proposing a framework that supports small-scale knowledge acquisition and modeling, and further refines the granularity of subject knowledge and sentiment. 2) On the basis of LDA theory, a concept hierarchy analysis method is introduced to design and implement a topic-clustered concept lattice generation algorithm for review documents, which is helpful for mining sentiment of sparse short text data.

The aim of the approach is not to demonstrate the advantages of utilizing LDA for opinion mining, but to construct a set of feature categories of students' comment for the online learning community. The innovative point proposed is to use the relationship between topic features and review text classification to build the formal context of formal concept analysis, thereby constructing a hierarchical topic concept lattice to reveal the implied sentiment.

For the sentiment classification of texts in online learning communities, review topics of students often have characteristics such as limitedness, which can lead to calculation errors in the sentiment similarity of reviews. In order to reduce the interference of topic content on sentiment classification, this article first mines the implicit topics of online reviews based on the LDA topic model, and combines the sentiment dictionary to calculate the sentiment polarity of the topics to obtain the sentiment tendency of the comments. To enact these goals, two approaches are proposed: on the one hand, the topics are detected via the LDA probability topic model; on the other hand, the sentiment scores matrix based on FCA is obtained by calculating sentiment similarities.

4 Topic sentiment analysis method

4.1 Design of the framework

The visualization for topic sentiment analysis in online learning community (TSAOLC) consists of four modules: data preprocessing, topic detection, sentiment analysis, and visualization. The architectural overview of TSAOLC is shown in Figure 2. The process of analyzing topic sentiment depends on a sequence of each step, which is depicted as follows.

Data preprocessing. Most of the text in OLC is unstructured or semi-structured text. Therefore, before extracting the text topic, the original text needs to be structured and represented in three steps: First, we use the web crawler tool to download related webpages and data sets; Then, stop word removal and Chinese word segmentation are applied to process the original corpora, which are used to edit the dictionary to get the document-word segmentation matrix; Finally, based

on the total frequency of occurrences, the top n terms for topic detection are selected to simplify relevant sets of candidate terms.

Topic detection. Firstly, the LDA probability topic model is used to obtain the subject candidate set with probability dependence in the document, during which the documents-topics probability matrix and the topics-terms probability matrix are obtained. The documents-topics probability matrix can be utilized to assign the membership values of topics to the correspondent relevant documents. Meanwhile, the top terms in the topics-terms probability matrix are likely to be semantically related, which can be clustered to form a certain meaningful concept. Therefore, the generated documents-topics probability matrix is regarded as the input for the formal context to construct the topic-clustered concept lattice.

Sentiment analysis. In the process of sentiment analysis, the common process is to extract the sentiment terms from the domain context and calculate the sentiment polarity with the help of the sentiment dictionary. In general, the main steps can be divided into two stages: 1) First, it is necessary to establish a sentiment vocabulary related to OLC; 2) Secondly, the sentiment scores are calculated to determine the semantic distance between the sentiment words.

Visualization. The process of visualizing topic sentiment can be divided into three stages. First, based on the FCA theory, a set of formal concepts containing topic-terms as well as a set of association rules are generated. Then, the sentiment scores matrix is mapped to the set of association rules to obtain a set of topic sentiment that satisfy different confidence levels. Finally, the negative sentiment is screened out as specific constraints to generate a sentiment rule set that reflects the negative tendency of the students in the text.

4.2 Data preprocessing

First, we set the website list of OLC to the seed URL, and use the web crawler software to download the web pages and data sets; then, after downloading the web pages to the local disk, the source codes are analyzed to extract the useful information containing the web page titles, which are saved in the database. Meanwhile, in order to filter irrelevant webpages, the unnecessary symbols are removed on the basis of the common stop words list belonging to a predefined domain vocabulary, which consists of Baidu stopwords vocabulary and machine learning stopwords list of Sichuan University. Afterwards, we utilize the Chinese word segmentation system named NLPPIR-ICTCLAS to process the text corpus, which can automatically discover new terms and adaptively test the linguistic probability distribution from longer text content. After extracting the candidate terms, TF-IDF (Term Frequency-Inverse Document Frequency) is introduced to assess the importance degree of a term to a document (Wu et al., 2008).

Figure 2. The overall structure of the proposed methodology.

4.3 Topic detection

Topic-modeling. The documents-terms matrix in Section 4.2 is used as the training initial document for the LDA model. The model can be divided into three steps: 1) Model initialization. For each document, a topic is extracted from the topic distribution. Meanwhile, a term is extracted from the term distribution corresponding to the extracted topic. The above process is repeated until each term in the document is traversed. 2) Model training. The training parameter set of the LDA model includes (α, β, k, n, i) , wherein the meaning of each parameter is explained as follows: α represents a symmetric Dirichlet parameter, which refers to the smoothness degree of the generated topic words; β represents an asymmetric Dirichlet parameter, which refers to the smoothness degree of topic feature to generate feature terms; k indicates the number of topics clustering; n indicates the number of feature terms under each topic; i indicates the number of iterations of the Gibbs sampling algorithm. In this paper, the values of the parameter sets (α, β, k, n, i) are set to $(1, 0.01, 50, 500, 5000)$. 3) Identity array definition. The Topics-Terms

matrix and the documents-topics matrix are generated after the edge distribution of these variables is obtained. We select the terms of top N according to the probability value ranking under each topic, and give an array of identifiers under different topics by combining the domain knowledge of OLC.

Formal concept analysis. In our case, we employ the FCA theory, as presented in section 3.2, to construct a hierarchical topic-clustered concept lattice, where the topics-terms matrix is fed as an input to be converted into a topic formal context represented as $K=(Documents, Topics, I)$. In the classical theory of FCA, the formal concept lattice can only be constructed through binary relations. However, the association matrix of the previous step usually contains multi-valued attributes that cannot be used to construct a topic clustering concept lattice directly. Therefore, we first use the conceptual scaling technique to structure and single-value the multi-valued attributes in the correlation matrix, and then convert the obtained multi-valued formal context into a formal context, which is named association matrix binarization. After that, the generated documents-topics matrix is assigned to the formal context to estimate the probabilities of each topic. Based on the analysis above, a proposed method algorithm for constructing the hierarchical topic-clustered concept lattice is summarized as shown in Table 1.

4.4 Sentiment analysis

Sentiment identification. In order to compute sentiment scores, a sentiment scores matrix (SSM) is created for storing the distance of semantic similarity among the opinion terms. We first extract the topic formal concept from the concept lattice in Section 4.3 to build a set of term datasets. In each dataset, documents and topics are viewed as rows and columns of the matrix, respectively. Then in order to obtain the sentiment polarity of the term, it is necessary to calculate the semantic similarity between the new term and the positive or negative seed terms on the basis of the sentiment dictionary, the results of which can be used as the basis for sentiment classification of the term. The construction of the sentiment dictionary consists of two steps, namely first assigning an initialized value to the seed term, and then assigning sentiment weights to the modifiers of the sentiment terms (such as degree adverbs, negative words, etc.). The collection of seed terms in the sentiment dictionary is generally divided into positive sentiment, negative sentiment and neutral sentiment. It is the basis for constructing a collection of sentiment words, which is mainly derived from the value of domain experts. After obtaining the seed terminology, we obtain the classification threshold of the sentiment polarity by calculating its mutual information. Then, by calculating the mutual information between the newly added term t_i and the known sentiment term t_j , the sentiment distance between them is obtained as shown in Equation 4. Modifiers such as

Research Paper

turning conjunctions, negative words, and degree adverbs in the sentiment dictionary often have volatility effects on the strength of sentiment words. Existing methods cannot effectively solve the problem of sentiment ablation when multiple fusions are performed. To enact this need, this article improves the above problems in two ways. On the one hand, before calculating the weight of the newly added term, we utilize the conditional clause after the transitional conjunction in the sentence to replace the whole sentence. At the same time, we calculate the sentiment comprehensive value based on the multi-feature linear fusion method to avoid the sentiment discrimination effectively. On the other hand, in order to quantify the influence of degree adverbs on sentiment intensity, we give the weights of common degree adverbs by referring to the literature (Gao, Luo, & Wang, 2017), namely “ adv_1 (1.5), adv_2 (1.3), adv_3 (1.1), and adv_4 (1). The synonyms of degree adverbs are also considered to have equal sentiment intensity, which are listed in Table 2. The Equation for calculating the sentiment comprehensive value is shown in Equation 5 on the basis of the reference (Wang, Pan, & Yang, 2019).

$$Sim_{MI}(t_i, t_j) = \sum_{t_i \in \{0,1\}} \sum_{t_j \in \{0,1\}} \log \frac{p(t_i, t_j)}{p(t_i)p(t_j)} \times p(t_i, t_j) \quad (4)$$

where $p(t_i, t_j)$ represents the probability that the terms t_i and t_j appear in the same document; $p(t_i)$ and $p(t_j)$ represent the probabilities that the document contains t_i and t_j respectively. The probabilities above-mentioned can be improved based on the sentiment dictionary proposed by the reference (Zhao et al., 2016) in the initial corpus using the method of maximum likelihood estimation.

$$SD(i_{new}, i_{seed}) = (adv_i) * (neg) * \left\{ \frac{\sum_{i_{seed} \in \{0,m\}} \sum_{i_{seed} \in S_{pos}} Sim_{MI}(i_{new}, i_{seed})}{|S_{pos}|} + \frac{\sum_{i_{seed} \in \{0,n\}} \sum_{i_{seed} \in S_{neg}} Sim_{MI}(i_{new}, i_{seed})}{S_{neg}} \right\} \quad (5)$$

where i_{new} represents the new term. i_{seed} represents the positive or negative seed terms. (adv_i) represents the weights for adverb of degree and its value range is 1, 1.1, 1.3, 1.5. (neg) is a negative variable in the sentence where i_{new} is located, and its value range is 1 or -1. When $(neg)=1$, it means that there is no negative term on the right side of i_{new} , and the sentiment tendency has not changed. When $(neg)=-1$, it indicates that there is a negative term on the right side of i_{new} , and the sentiment tendency are opposite to the existing sentiment words in this sentence. S_{pos} and S_{neg} represent the set of positive sentiment terms and negative sentiment terms, respectively, m and n represent the number of terms contained in the set of the positive sentiment terms and the negative sentiment terms, respectively. If $SD(i_{new}, i_{seed})$ is greater than 0, it means that i_{new} has a positive tendency and should be added into the set of the positive sentiment terms and vice versa.

Sentiment score calculation. In order to obtain a student's sentiment tendency towards a topic, the SSM needs to be initialized. The main process can be described as follows. Firstly, the student's comment topics are extracted from the topic formal context to establish the text-topic matrix. Secondly, the similarity distance between different topics is calculated to obtain the mapping relationship between students and topics. Afterwards, the maximum sentiment value of the sentiment terms under different topics is calculated to get the sentiment similarity distance that converges under a certain topic.

Topic similarity computing. KL distance is utilized to calculate the similarity distance under different topics—terms probability distribution (Zheng et al., 2013). However, the KL distance is asymmetrical, which is unable to compute the symmetric topic distribution functions. Therefore, enlightened by the literature (Wang, Zuo, & Tao, 2015), we introduce the relative entropy into the topic similarity calculation to iterate over all the terms, which is computed using Equation 6.

$$sim_{KL}(t_i, u_j) = -\frac{1}{2} \left(\sum_{t_i \in (0,1), u_j \in (0,1)} p(t_i) \log \frac{t_i}{u_j} + \sum_{t_i \in (0,1), u_j \in (0,1)} q(u_j) \log \frac{u_j}{t_i} \right) \quad (6)$$

where (t_i, u_j) represents two topics; $p(t_i)$ and $q(u_j)$ represent the distribution function of the topics under their respective conditional probabilities.

Sentiment similarity computing. After obtaining the mapping relationships between the student and the topic, it is possible to calculate the sentiment value (as shown in Equation 7) of the student and obtain the sentiment tendency under a certain topic.

$$Sentiment_{score}(i_n^{d,t}) = \sum_{j=1}^d \sum_{k=1}^t SD(i_{new}, i_{seed}) \quad (7)$$

Here, d and t are the sizes of documents and topics, respectively; SD represents the sentiment comprehensive value.

On the basis of the equations above-mentioned, the proposed algorithm in Table 3 for calculating SSM is listed as follows: Step 1 to Step 3 initialize the sentiment score and the probability distribution $p(t_i, t_j)$ as 0. Step 4 derives the positive and negative seed terms to construct a collection of sentiment words. Step 5 computes the sentiment comprehensive value between each topic and seed terms. Step 6 computes the topic similarity between each topic of the students and each topic in the topic formal context. Step 7 to Step 13 iterates over all terms contained in the topic formal context. Step 8 initializes the sentiment comprehensive value. Afterwards, Steps 9 to 11 computes the sum of the membership values. Step 12 computes the whole sentiment scores of each topic of the students.

Research Paper

Table 1. A proposed method algorithm for topic-clustered concept lattice generation.

Input: A set of topic and comment documentation D , where $ D =n$, the number of potential topics m .	
Output: A topic-clustered concept lattice CL , a topics-terms probability matrix P and a documents-topics probability matrix R .	
1.	for each $d_i \in D$.
2.	$d_i \rightarrow CWS_i$. // Convert the document into a word segment.
3.	for each cws in CWS_i .
4.	$W = W \cup \{cws\}$. // Obtain a collection of phrases that contains topic attribute.
5.	end for.
6.	end for.
7.	for each cws in CWS_i .
8.	$CWS_i \rightarrow tfidf_i$. // Calculate the term frequency of attributes.
9.	$\dot{D} = [D : tfidf_i]$. // Obtain term frequency vector.
10.	end for.
11.	$(\dot{D}, W) \xrightarrow{LDA} (D, P, I)$. // Perform topic detection.
12.	$(\dot{D}, P) \rightarrow R$. // Classify topic association matrix.
13.	Find the subset of topic attributes represented as t_j .
14.	for $j=1$ to 2^m .
15.	Compute the set of objects by applying the Glöis connection.
16.	$R \rightarrow I'$. // Convert topic association matrix to multi-valued formalcontext.
17.	$I' \rightarrow I$. // Convert multi-valued formalcontext to binary single-valued formalcontext.
18.	$(\dot{D}, R, I') \xrightarrow{FCA} CL(D, R, I')$. // Construct a hierarchical topic concept lattice.
19.	end for.
20.	Return $\{CL(\dot{D}, R, I'), P, R\}$.
21.	Derive the topic-clustered sets.

Table 2. Classification weights for adverb of degree.

Level(weights)	Included adverbs
$adv_1(1.5)$	excessively, completely, extensively, dreadfully, entirely, absolutely
$adv_2(1.3)$	fairly, pretty, rather, quite, very, much, greatly, by far, highly, deeply
$adv_3(1.1)$	really, almost, nearly, bven, just, still
$adv_4(1)$	slightly, a little, a bit, trifle, somewhat

Sentiment adverbs in Table 2 is to play the role of an adverb modifying the whole sentence, which can be a more accurate understanding of the performance of those comments, thoughts and experiences in the comment text. Besides, the stronger the sentiment polarity expressed are, the more in line with students demand for learning effects to understand and analyze. Therefore, it is very necessary to quantify the influence of degree adverbs on sentiment intensity. We classify sentiment adverbs into 4 levels to obtain common degree adverbs on the basis of the literature (Zhang et al., 2017).

4.5 Visualization based on FCA

After obtaining SSM, the generated topic set of formal concepts in the previous section is regarded as an input to ConExp 1.3, which can output two association rules under different confidence values: weak association rules and strong association

Table 3. A proposed method algorithm for calculating sentiment scores matrix.

Input: A topic formal context $K=(U, T, I)$, where $U=\{u_1, u_2, \dots, u_n\}$ represents a set of topics belonging to a group of students, $T=\{t_1, t_2, \dots, t_m\}$, n is the size of student set, m is the size of topics.

Output: A sentiment scores matrix $Sentiment_{score}(t_i)$, where i represents sentiment score of each topic.

1. for each topic t_i in T .
2. $Sentiment_{score}(t_i)=0$.
3. $P(t_i, u_i)=0$.
4. Derive the positive and negative seed terms on the basis of domain experts.
5. Compute $sim_{KL}(t_i, u_j)$ // Compute the mutual information.
6. Compute $SD(t_i, t_{seed})$ // Compute the sentiment comprehensive value.
7. for each topic of student u_j in the topic formal context K .
8. $SD(t_i, u_j)=0$.
9. for each topic of t_i in the topic formal context K .
10. $SD(t_i, u_j)=SD(t_i, t_{seed})+SD(u_j, t_{seed})$.
11. end for.
12. $Sentiment_{score}(t_i)=Sentiment_{score}(t_i)+SD(t_i, u_j)$.
13. end for.
14. end for.
15. Return $Sentiment_{score}(t_i)$.

Note: For the selection of positive and negative seed terms, domain expert refers to ten participants, including authors, who use Borda counts to vote on different seed terms. Specifically, for any term to be classified, it is called three alternative sentiment datasets (positive sentiment, negative sentiment, and neutral sentiment), sorted by score, and finally classified as the highest according to the majority voting principle.

rules. The weak association rule set is also named Luxenburger set of approximate rules, while strong association rules are called Duquenne-guigues set of implication rules where the degree of support values and confidence values are both greater than the minimum support and the minimum confidence threshold (Qodmanan, Nasiri M., & Minaei, 2011). In order to analyze the sentiment state of online learning in a more in-depth way, we utilize the both association rules to map each topic to its relevant sentiment. To enact this need, the weak association rules are generated by using “Calculate Duquenne-guigues set of implications” module to create implication rules. Meanwhile, the strong association rules are also generated by using “Calculate Association Rule” module. The generated association rule expression is of the form “*Number of objects* < *Number of objects satisfying the preconditions* > *Precondition* = [*Confidence*] => < *Number of objects satisfying the preconditions and conclusions* > *Conclusion*”. The most effective association rules are obtained by adjusting the number of association rules along with the minimum confidence and minimum support.

The same student has different information needs in different situations, so the relative identities and needs of students in OLC are dynamic. Therefore, the current method focuses on mining association rules of learning identities to identify the transformational rules of students' relative roles in different situations, which helps to realize the precise service of the community. In addition, mining the association

Research Paper

rules for student behaviors can help to establish the sentiment evolution path of students on specific topics, which can improve the basis for the transformation of roles from different learning groups. Therefore, we also discuss the behavioral association rules. A detailed explanation of the two association rules will be explained in the Section 6.

5 Implementation

In this part, we develop the model of topic sentiment analysis in OLC on several modules to express and monitor opinions. The first module for data preprocessing collects 171,430 comments by crawling the text corpora from <http://www.icourses.cn/home/> on four categories namely Computer science, Economic management, Medicine & Health, and Agriculture and forestry. The second module is used for detecting discussion topics to construct hierarchical topic-clustered concept lattice. The topic terms and their corresponding probability distributions are shown in Table 4, where each topic is represented by the top 5 terms with the highest probability. The third module is designed for computing SSM in order to obtain the sentiment information hidden in the topics. The topic formal context is on the basis of the multi-valued formal context with the sentiment scores of each topic as shown in Table 5. Afterwards, we convert the formal context in Table 5 into a binary single-valued formal context, via the association matrix binarization based on positive, objective, and negative relationships. However, since information in Table 5 does not relate to the sentiment of the students during the learning process, it is relatively difficult to analyze the association rules between the student’s identity and the learning behavior. To enact this need, we further mine the terms from the relevant documents containing student identity information and learning behaviors. Specifically, the student identity attributes mainly contain “*learner*”, “*administrator*”, “*freshman*”, “*junior*”, “*postgraduate*”, “*psychological stress*”. The learning behavior attributes mainly contain “*cooperation*”, “*interaction*”, “*information provider>AVG*”, “*information provider<AVG*”, “*information searcher>AVG*”, “*information searcher<AVG*”, “*information sharer>AVG*” and “*information sharer<AVG*”. Therefore, we add the attributes above-mentioned (columns of Table 6) to get the sentiment information of the students during the learning process, and the binary sentiment of the single-valued formal context is shown in Table 6. The fourth module shows a flat structure view of the associated topics-sentiment by mapping topics to their association rules.

Figure 3 represents a screenshot of the proposed method, which enables teachers and supervisors to view each specific topic by adjusting the controls of the browser. Supervisors can select each topic (top-right section) based on LDA with its most

relevant topics-terms matrix (middle-right section) and documents-topics matrix (middle-left section). Besides, five documents are randomly selected to be assigned to the topic whose TF-IDF values are greater than the pre-setting threshold (bottom-right section). Afterwards, the hierarchical topic concept lattice is constructed based on FCA (central part). Finally, two sets of the implication rules and association rules are listed on the basis of the sentiment scores matrix (bottom-left part).

Table 4. Recognition results of topic terms.

Topic	Term and its probability
T ₁	Course selection/0.023, Learning objectives/0.021, Difficulty of knowledge/0.018, Teaching methods/0.017, Guidance methods/0.013
T ₂	Credits/0.025, Content organization/0.023, Teaching methods/0.021, Learning support/0.021, Homework and assessment methods/0.020
T ₃	Case presentation/0.032, Procedural evaluation/0.031, Knowledge expansion/0.029, Analysis of difficult points/0.027, Group discussion/0.027
T ₄	Communication and feedback/0.033, Resource sharing/0.033, Information update/0.032, Response time/0.031, Information acceptance/0.030

T₁: Instructional design; T₂: Course content; T₃: Teaching effect; T₄: Teaching interaction.

Table 5. Multi-valued sentiment formal context based on topic association matrix.

	T ₁	T ₂	T ₃	T ₄
D ₁	-3.427	2.874	4.315	-1.306
D ₂	2.641	-0.597	-2.105	2.635
D ₃	4.715	2.132	1.624	0
D ₄	2.334	0	-1.748	4.316
D ₅	-3.619	-1.857	3.624	-0.391
D ₆	-2.107	2.167	2.419	2.361
D ₇	0	-0.524	-0.267	2.638
D ₈	2.369	1.629	2.364	0
D ₉	1.024	-0.121	3.478	2.964
D ₁₀	2.361	1.493	-0.328	-1.267

The platform represented by Figure 3 is constructed based on two open source tools, namely the interactive visualization library pyLDAvis and the open-source tool Colibri / ML. On the one hand, for the analysis and discovery of topics, the topic model interactive visualization library pyLDAvis is introduced to semi-automatic mining of potential comment topics from unstructured text resources. The above process mainly includes preprocessing the data, generating a document word frequency matrix, generating an LDA model, and mining association rules. First, in the data preprocessing phase, for a large number of html tags, non-Chinese characters are eliminated. Secondly, in the frequency matrix generation stage, Chinese text analysis software is used to analyze the text data to meet the basic requirements of machine learning. At the same time, words that are meaningless are excluded.

Table 6. The binary sentiment of the single-valued formal context.

	T ₁	T ₂	T ₃	T ₄	T ₅	T ₆	T ₇	T ₈	T ₉	T ₁₀	T ₁₁	T ₁₂	T ₁₃	T ₁₄	T ₁₅	T ₁₆	T ₁₇	T ₁₈	T ₁₉	T ₂₀
D ₁	*			*			*	*		*			*				*			
D ₂			*		*	*		*	*	*		*		*	*			*		*
D ₃				*			*	*			*		*		*			*		
D ₄	*				*	*		*			*	*	*		*				*	
D ₅		*			*	*		*	*	*		*	*		*		*			*
D ₆		*			*	*		*			*	*	*		*		*			
D ₇			*		*		*	*			*	*	*	*	*	*		*		*
D ₈	*				*		*	*			*	*	*	*	*	*		*		*
D ₉	*				*	*	*	*		*		*	*	*	*	*		*		*
D ₁₀			*		*	*	*	*		*		*	*	*	*	*	*	*		*

Note: * represents criterion satisfied, T₁ represents *freshman*, T₂ represents *junior*, T₃ represents *postgraduate*, T₄ represents *administrator*, T₅ represents *learner*, T₆ represents *information provider* <AVG; T₇ represents *information provider* >AVG; T₈ represents *information sharer* <AVG; T₉ represents *information sharer* >AVG; T₁₀ represents *information searcher* <AVG; T₁₁ represents *information searcher* >AVG; T₁₂ represents *psychological stress*, T₁₃ represents *cooperation*, T₁₄ represents *interaction*, T₁₅ represents P_{T11}, T₁₆ represents O_{T11}, T₁₇ represents N_{T11}, T₁₈ represents P_{T12}, T₁₉ represents O_{T12}, T₂₀ represents N_{T12}, T₂₁ represents P_{T13}, T₂₂ represents O_{T13}, T₂₃ represents N_{T13}, T₂₄ represents P_{T14}, T₂₅ represents O_{T14}, T₂₆ represents N_{T14}, * represents criterion satisfied. As the length limits, the topics of T₂₁ to T₂₆ is not shown in this table.

Figure 3. A screenshot of the tool of documents-topics for sentiment mining.

Finally, in the LDA topic generation stage, the topic number K needs to be set in advance, and then the LDA model is trained by adjusting model parameters to generate a model_twosds.dat file containing (topicID, word, probability). After that, different topics and related words can be obtained to realize the interactive display. In addition, pyLDavis can fine-tune lambda parameters to achieve dynamic adjustment of word weights. For the second part, FCA can build a concept lattice based on the “document-topic-probability” matrix output by LDA. An open-source tool Colibri / ML is introduced, which is developed to implement a novel geometric representation of programming structural patterns and violations to figure out patterns. Specifically, first, the probabilistic relationship between the subject concepts is mapped to the upper and lower semantic relationship between the object and the attribute respectively, and Colibri-java is used to perform formal context reduction, purification and other editing. Secondly, based on the topic formal context, the concept hierarchy analysis is performed on the “document-topic” matrix to generate the hierarchical topic concept lattice, and the upper-lower relationship

between topics is obtained. The result is displayed in the document tree view in the form of a Hasse diagram. Finally, we use the plugin “Calculate Association Rules” of ConEXP 1.3 to realize the mining of association rules implicit in the topical context.

6 Results and discussions

In order to verify the effectiveness and accuracy of the proposed model in mining learning behaviors and the hidden sentiment under different learning conditions, we first analyze the implication rules and association rules generated in the third module in Section 5. Afterwards, mean absolute error, precision, recall and F values are used to measure the overall performance of the proposed method, compared with other state-of-art models.

6.1 Analysis of the implication rules and association rules

The model generates a total of 164 association rules with a confidence level greater than 50%, and 97 implication rules. As the negative sentiment can express a stronger need of information on certain topics compared to the positive sentiment. Thus, a list of total 48 association rules and 47 implication rules are obtained. In addition, by adjusting the minimum confidence and minimum support, the most efficient of the rules above-mentioned can be highlighted under a smaller number of rule conditions. Let us assume that the support degree is three, and the confidence is greater than 50%, the remaining ten association rules are selected when the preconditions in the rule sets contain “learner” and the conclusions involve student behaviors. Similarly, when the preconditions in rule sets relate to learning behavior and the conclusion contains student identity information or student behavior, three implication rules are obtained. The selected rules are listed in Table 7. In the sentiment mining of association rules, three valuable basic rules can be summed up. Firstly, students who are learning online have a great possibility to meet the dynamic needs of learning by adjusting learning behaviors. Rule 1 indicates that students who do not like to provide information will obtain information through information tracking when they are dissatisfied with the content of the course. Secondly, when students with relatively higher academic levels express negative sentiments, they often improve their learning effects through communication with professors, academic authorities, etc. Rule 6 shows that when learners are psychologically stressed and dissatisfied with the learning effect, there is a 67% chance that the learner is a graduate student and they will be willing to communicate their information needs by interacting with others. Finally, when students are dissatisfied with a topic, they usually adjust to themselves through some inherent learning behavior habits,

which reflect the autonomy of students in OLC. Rule 4 shows that when learners are dissatisfied with the learning content, there is a 75% chance of dealing with related issues through information tracking and active sharing. Besides, the basic rules hidden in the implication rule sets can be summarized into two points. For the one hand, the attitude of the students to the teaching effect will greatly affect their learning status, and the learning behavior with stress will change accordingly. Rule 2 indicates that when students who tend to interact and are willing to share their personal attitudes show anxiety about the teaching effect, they are often reluctant to alleviate their psychological stress through information search. For another hand, there is a relatively strong correlation between different learning behaviors. Both Rules 1 and 3 indicate that students, who are likely to share information (retrieve information) and are willing to cooperate with each other, often track relevant information (share information) to conduct more in-depth learning and show more positive learning sentiment. This conclusion indicates that it is more important to cultivate students' good learning behavior habits compared with the course content, which is of great guiding significance for improving students' online learning efficiency.

6.2 Experimental verification and evaluation criteria

The experimental data of this paper is selected from the topics of "Computer Science", "Information Science", "Network Engineering" and "Software Engineering" in <http://www.icourses.cn/home/>. The time starts from January 7, 2019 to January 13, 2019, which is the last week before the final test in many schools. The number of active users online was 40,758, and the number of valid comments was 102,846. The text corpus is divided into 7 consecutive time segments, which are named St1-St7 in turn. Then, we input the data set into the following models to calculate topic sentiment values, namely the method proposed (TSAOLC), the association rule algorithm (RA) (Zhi, 2002), the clustering algorithm (CG) (Meng, Shen, & Chen, 2013), TextBlob (<https://github.com/sloria/TextBlob>), and the co-training algorithm (CoT) (Hady & Schwenker, 2008). The results of each method are ranked according to sentiment values in their descending order and compared to the results verified by human annotator (the ranking bias is set to ± 6) (Chen, 2018).

Besides, to evaluate the advantages and effectiveness of the proposed method, we further performed a comparative analysis with the semi-supervised method (Co-Training algorithm) based on SVM Classifier on high-dimensional datasets, compared with Naïve Bayes, multilayer perceptron and random forest, to select a suitable classifier to build a predictive model for the quality of topic sentiment analysis.

Research Paper

The calculation results of relevant evaluation indicators are shown in Tables 8–10. The results show that TSAOLC exhibits high classification performance on all datasets, which validates the effectiveness and stability of the proposed method.

In order to select a suitable classifier to establish a quality prediction model, we further performed a comparative analysis with Naïve Bayes, multilayer perceptron and random forest. The maximum number of decision trees in a random forest is 100. The hidden layer of the multilayer perceptron is 3, and the learning rate is 0.2.

The evaluation performance of partial data in different classifiers is shown in Figures 4–5. From the experimental results in Fig 4, it can be known that the precision, recall and F-measure of the topic sentiment prediction model of all the data can be around 0.5. Meanwhile, the value of MAE stays below 0.7. Based on the TSAOLC model defined in this paper, the comprehensive performance using the random forest method is optimal, which means that the weighted average of various indicators on the three sub-data sets is the best (Average MAE is 0.4758).

Table 7. The implication rules and association rules.

Association rules	<p>1<3>Learner Information provider<AVG NT2=[100%]>=><3>Information searcher>AVG; 2<4>Learner Psychological stress PT1=[75%]>=><3>Information provider<AVG NT3; 3<4>Learner NT2=[75%]>=><3>Interaction; 4<4>Learner NT2=[75%]>=><3>Information sharer>AVG Information searcher>AVG; 5<3>Learner Information sharer<AVG Psychological stress Cooperation PT1=[67%]>=><2>Information provider<AVG NT3; 6<3>Learner Information searcher>AVG Psychological stress PT1 NT3=[67%]>=><2>Postgraduate Information searcher<AVG Interaction; 7<3>Learner Information provider<AVG PT1 PT4=[67%]>=><2>Information searcher<AVG NT2; 8<3>Learner Information provider<AVG Information searcher<AVG NT2=[67%]>=><2>Information sharer>AVG Psychological stress Interaction; 9<3>Learner NT2 PT4=[67%]>=><2>Postgraduate Interaction; 10<3>Learner NT2 PT4=[67%]>=><2>Information sharer<AVG;</p>
Implication rules	<p>1<2>Learner Information sharer>AVG Interaction cooperation ==> Information searcher>AVG Psychological stress PT2; 2<2>Learner Interaction sharer>AVG NT3==> Information searcher<AVG Psychological stress; 3<2>Learner Information searcher>AVG Interaction cooperation ==> Information sharer>AVG Psychological stress PT4;</p>

Note: The pre-setting condition for the association rule is (Preconditions contain learners= [>50%] => Conclusions related to student behavior); The pre-setting condition for the implication rule is (Preconditions related to student behaviors => Conclusions related to student identities or student behaviors). When the frequency of the user behavior in Table 7 is greater than the mean value, it can be considered that under the constraint of the precondition, the student has a relatively high probability to adopt such behavior.

6.3 Illustrative example

To better illustrate how the proposed method can help teachers or supervisors implement teaching process management, the construction for topic-clustered concept lattice can be divided into description layer, topic feature layer, learning

Table 8. Precision contrast between different methods based on SVM.

	S ₁₁	S ₁₂	S ₁₃	S ₁₄	S ₁₅	S ₁₆	S ₁₇
RA	49.32	37.51	40.67	42.52	43.77	41.26	45.33
CG	52.33	34.96	38.79	41.68	40.17	37.74	42.59
CoT	57.73	46.28	48.85	44.84	51.39	47.77	48.25
TextBlob	58.86	45.16	46.07	42.33	52.78	45.56	52.63
TSAOLC	61.34	50.23	54.95	49.83	53.95	62.98	54.36

Table 9. Recall contrast between different methods based on SVM.

	S ₁₁	S ₁₂	S ₁₃	S ₁₄	S ₁₅	S ₁₆	S ₁₇
RA	44.45	42.06	47.64	44.37	45.98	41.63	48.21
CG	42.68	40.97	48.86	42.07	43.63	42.88	47.71
CoT	49.99	47.38	52.84	55.36	52.09	49.23	53.84
TextBlob	54.18	45.84	51.67	58.07	62.29	53.46	60.06
TSAOLC	56.49	58.03	62.27	59.96	65.59	58.76	62.34

Table 10. F-measure contrast between different methods based on SVM.

	S ₁₁	S ₁₂	S ₁₃	S ₁₄	S ₁₅	S ₁₆	S ₁₇
RA	46.67	39.65	43.88	43.43	44.85	41.44	46.73
CG	47.01	37.73	43.25	41.87	41.83	40.15	45.00
CoT	53.58	46.82	50.77	49.55	51.74	48.49	50.89
TextBlob	56.42	45.50	48.71	48.97	57.14	49.19	56.10
TSAOLC	58.82	53.85	59.38	54.43	59.20	60.80	58.08

Table 11. MAE contrast between different methods based on SVM.

	S ₁₁	S ₁₂	S ₁₃	S ₁₄	S ₁₅	S ₁₆	S ₁₇
RA	98.42	92.46	90.87	88.38	89.07	91.45	95.63
CG	82.03	85.56	87.69	89.06	92.61	94.97	86.36
CoT	78.84	76.34	72.19	68.78	75.43	76.35	78.62
TextBlob	72.93	67.45	69.37	64.92	70.14	68.62	62.15
TSAOLC	58.99	54.56	57.32	55.25	57.20	59.15	53.13

sentiment analysis layer and visualization layer. The description layer uses data preprocessing technology to mine document-text matrix from student text, which mainly includes text content such as community postings, classroom discussions and students' online course selection. The classroom topic feature layer and learning sentiment analysis layer provide statistical information on learning topic content and subject-level clustering, so that teachers and managers can view each specific topic by adjusting the controls of the browser. Specifically, teachers and supervisors can select specific topic words to obtain a vector of topic variables and visualize hierarchical dependencies between topics. The visualization layer mainly includes the dynamic display of hierarchical topic-clustered concept lattice and the visualization of association rules. In the constructed concept lattice, super-concepts

Figure 4. Precision (left) and recall (right) comparison based on various classifiers.

Figure 5. F-measure (left) and MAE (right) comparison based on various classifiers.

have more extensions than sub-concepts, and sub-concepts have richer connotations than super-concepts. Among them, a white semicircle node indicates that the concept has an attribute, and a black semicircle node indicates that the concept has an object. As the level increases, the attributes of the layer concept gradually increase, the number of objects with these concepts gradually decreases, and finally a specific object is located. Teachers can obtain all formal concepts that contain the topic word by selecting appropriate confidence thresholds. At the same time, they can use the attributes as prerequisites in association rule sets to obtain association rules and replication rules, so as to identify the student group’s sentimental tendency on specific topics. In addition, teachers or teaching managers can dynamically display topic concepts with the same clustering characteristics by clicking on different

concept nodes of the concept lattice. And if the topic concept set above-mentioned is used as a prerequisite of association rules, and the number of association rules is adjusted by adjusting the minimum confidence and the minimum support degree, a more concentrated negative sentiment evaluation in a certain type of topics can be obtained, thereby providing a reasonable basis for curriculum reform.

7 Conclusions and perspectives

This paper designs a model for online sentiment analysis of various topics in OLC. The model obtains the topic-terminology hybrid matrix and the document-topic hybrid matrix by selecting the real user's comment information on the basis of LDA topic detection approach. Afterwards, a topic clustering concept lattice based on FCA model is constructed, where the topic sentiment can be identified by measuring their sentiment scores. In addition, the topic sentiment can be visualized based on the implication and association rules to refine the granularity of sentimented knowledge. Finally, from the results of the experiment, the following conclusions can be obtained:

- The proposed model can effectively perceive students' sentiment tendencies on different topics, which provides powerful practical reference for improving the quality of information services in teaching practice.
- The topic-sentiment visualization framework can clarify the hierarchical dependencies between different topics, which lay the foundation for improving the accuracy of teaching content recommendation and optimizing the knowledge coherence of related courses.

In order to improve the accuracy of the topic-sentiment analysis model, the follow-up research will focus on optimization of semantic constraint capabilities between different topics. In addition, exploring the intensity of students' sentiments and their evolutionary trends will also be an interesting content, which will improve the adaptive ability of opinion mining.

Competing interests

The authors have declared that no competing interests exist.

Availability of data and material

All data generated or analyzed during this study are included in this article. If the reviewer is interested in the experimental data of this article, you can directly contact the author to request procedural verification data, please contact email: wangkai0552@126.com.

Acknowledgment

The authors would like to thank all the anonymous reviewers and the Editor-in-Chief for their valuable time to improve our paper. This work was partly supported by the Teaching Research Major Projects of Anhui Province (2018jyxm1446), the Natural Scientific Project of Anhui Provincial Department of Education (KJ2019A0371), the Anhui Demonstration Experiment Training Center Project (2018sxzx58) and the Demonstration Projects for Massive Open Online Course of Anhui Province (2018mooc278).

Author contributions

Kai Wang (wangkai0552@126.com) designed the research and wrote the paper. Yu Zhang (475870652@qq.com) analyzed data and performed the research.

References

- Kaklauskas, A., Zavadskas, E.K., Seniut, M., et al. (2013). Recommender system to analyze student's academic performance. *Expert Systems with Applications an International Journal*, 40(15), 6150–6165.
- Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(9), 993–1022.
- Cho, M.H., Kim, Y., & Choi, D.H. (2017). The effect of self-regulated learning on college students' perceptions of community of inquiry and affective outcomes in online learning. *Internet & Higher Education*, 34, 10–17.
- Chu, K.M., & Li, F. (2010). Topic evolution based on LDA and topic association. *J. Shanghai Jiao tong Univ. (Sci)*, 44(11), 1501–1506.
- Cerulo, L., & Distant, D. (2013). Topic-driven semi-automatic reorganization of online discussion forums: A case study in an e-learning context. *Global Engineering Education Conference. IEEE*, March 13–15, Berlin, Germany. pp 289–298.
- Chen, Y.C. (2018). A novel algorithm for mining opinion leaders in social networks. *World Wide Web-internet & Web Information Systems*, 2, 1–17.
- Ethem, F.C., Aysu, E.C., & Fazli, C. (2018). Multilingual sentiment analysis: An RNN-based framework for limited data. In *Proceedings of ACM SIGIR 2018 Workshop on Learning from Limited or Noisy Data*, July 12, Michigan, USA, pp 1–5.
- Fariza, K. (2019). Students' identities and its relationships with their engagement in an Online Learning Community. *International Journal of Emerging Technologies in Learning*, 14(5), 4–19.
- Colace, F., De Santo, M., & Greco, L. (2014). SAFE: A sentiment analysis framework for e-Learning. *International Journal Of Emerging Technologies In Learning (IJET)*, 9(6), pp. 37–41. doi:<http://dx.doi.org/10.3991/ijet.v9i6.4110>
- Ficamos, P., Yan, L., & Chen, W. (2013). A naive bayes and maximum entropy approach to sentiment analysis: Capturing domain-specific data in Weibo. *IEEE International Conference on Big Data & Smart Computing*. February 13–16, Jeju, South Korea. pp 324–332.

- Gao, G., Luo J.M., & Wang, Y. (2017). Analyzing textual sentiment based on HNC theory. *Data Analysis and Knowledge Discovery*, 8(8), 85–90.
- Ge, G., Chen, L., & Du, J. (2013). The research on topic detection of microblog based on TC-LDA. *IEEE International Conference on Communication Technology*. November 17–19, Guilin, China. pp 257–262.
- Ghiasifard, S., Khadivi, S., Asadpour, M., et al. (2015). Improving the quality of overlapping community detection through link addition based on topic similarity. *International symposium on Artificial Intelligence & Signal Processing (AISP 2015)*, March 3–5, Mashhad, Iran. pp 244–250.
- Hofmann, T. (1999). Probabilistic latent semantic indexing. *International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp 50–57.
- Kohoulat, N., Hayat, A.A., Dehghani, M.R., Kojuri, J., & Amini, M. (2017). Medical students' academic emotions: the role of perceived learning environment. *Journal of Advances in Medical Education & Professionalism*, 5(2), 78–83.
- Li, X.D., Ba, Z.C., & Huang, L. (2015). A text feature selection method based on weighted latent dirichlet allocation and multi-granularity. *New Technology of Library and Information Service*, 258, 42–49.
- Li, Y., Ma, S., Zhang, Y, Huang, R, & Kinshuk. (2013). An improved mix framework for opinion leader identification in online learning communities. *Knowledge-Based Systems*, 43(2), 43–51.
- Liu, Y., Bi, J.W., & Fan, Z.P. (2017). A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm. *Information Sciences*, 394, 38–52.
- Lu, Y., Zhang, P., Liu, J., Li, J., & Deng, S. (2013). Health-related hot topic detection in online communities using text clustering. *PLoS ONE*, 8(2), e56221.
- Martíneztorres, M.R. (2015). Content analysis of open innovation communities using latent semantic indexing. *Technology Analysis & Strategic Management*, 27(7), 859–875.
- Meng, Z.Q., Shen, S.M., & Chen, Q.L. (2013). A network decomposition-based text clustering algorithm for topic detection. *Applied Mechanics & Materials*, 239, 1318–1323.
- Mertiya, M., & Singh, A. (2016). Combining naive bayes and adjective analysis for sentiment detection on Twitter. *2016 IEEE International Conference on Inventive Computation Technologies (ICICT)*. August 26–27, Coimbatore, India. pp 157–163.
- Cheng, M.M., Su, C.Y., Zhang, J.P., & Yang, Y. (2015). Analyzing temporal patterns of groups and individuals in an online learning forum. *International Journal of Emerging Technologies in Learning*, 10(5), 66–71.
- Hady, M.F.A., & Schwenker, F. (2008). Co-training by Committee: A New Semi-supervised Learning Framework. *Workshops IEEE International Conference on Data Mining*. IEEE.
- Nagori, R., & Aghila, G. (2012). LDA based integrated document recommendation model for e-learning systems, *IEEE International Conference on Emerging Trends in Networks & Computer Communications*. April 22–24, Udaipur, INDIA. pp 204–215.
- Nan, L., & Wu, D.D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, 48(2), 354–368.

Research Paper

- Qodmanan, H.R., Nasiri, M., & Minaei-Bidgoli, B. (2011). Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence. *Expert Systems with Applications*, 38(1), 288–298.
- Parvathy, G., & Bindhu, J.S. (2016). A probabilistic generative model for mining cybercriminal network from online social media: A review. *International Journal of Computer Applications*, 134(14), 1–4.
- Pappas, N., Redi, M., Topkara, M., et al. (2017). Multilingual visual sentiment concept clustering and analysis. *International Journal of Multimedia Information Retrieval*, 6(1), 51–70.
- Ren, R., Ling, W., & Yao, Y. (2018). An analysis of three types of partially-known formal concepts. *International Journal of Machine Learning & Cybernetics*, 9(11), 1767–1783.
- Santosh, D.T., Vardhan, B.V., & Ramesh, D. (2016). Extracting product features from reviews using feature ontology tree applied on LDA topic clusters. *IEEE 6th International Conference on Advanced Computing (IACC)*, February 27–28, Bhimavaram, India. pp 89–96.
- Shirakawa, M., Nakayama, K., Hara, T., et al. (2017). Wikipedia-based semantic similarity measurements for noisy short texts using extended naive bayes. *IEEE Transactions on Emerging Topics in Computing*, 3(2), 205–219.
- Shin, B., Lee, T., & Choi, J.D. (2017). Lexicon integrated CNN models with attention for sentiment analysis. *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, September 7–11, Copenhagen, Denmark, pp 149–158.
- Shea, P., Li, C.S., & Pickett, A. (2006). A study of teaching presence and student sense of learning community in fully online and web-enhanced college courses. *Internet & Higher Education*, 9(3), 175–190.
- Turney, P.D., & Littman, M.L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4), 315–346.
- Vinodhini, G. (2014). Sentiment mining using SVM-based hybrid classification model. *Advances in Intelligent Systems & Computing*, 246, 155–162.
- Wang, J., Zuo, W., & Tao, P. (2015). Hyponymy graph model for word semantic similarity measurement. *Chinese Journal of Electronics*, 24(1), 96–101.
- Wang, K., Pan, W., & Yang, B.H. (2019). Analysis of topic emotion evolution based on OTSRM model. *Journal of Information*, 38(05), 534–542.
- Wei, L., Wang, Z., Qian, T., & Wan Q. (2019). Attribute reduction in the background of multi-source decision forms. *Journal of Shaanxi Normal University (Natural Science Edition)*, 47(5), 57–63.
- Wu, H.C., Luk, R.W.P., Wong, K.F., & Kwok, K.L. (2008). Interpreting TF-IDF term weights as making relevance decisions. *Acm Transactions on Information Systems*, 26(3), 55–59.
- Wu, W.C.V., Hsieh, J.S.C., & Yang, J.C. (2017). Creating an online learning community in a flipped classroom to enhance EFL learners' oral proficiency. *Journal of Educational Technology & Society*, 20(2), 142–157.
- Xie, X., Ge, S., Hu, F., Xie, M., & Jiang, N. (2017). An improved algorithm for sentiment analysis based on maximum entropy. *Soft Computing*, 23(1), 599–611.
- Yang, M., Peng, B.L., & Chen, Z. (2014). A topic model for building fine-grained domain-specific emotion lexicon. *Association for Computational Linguistics (ACL)*, pp 421–426.

- Yang, C., Zhang, H., & Shi, D. (2014). An on-line adaptive topic evolution model in web discussions. *IEEE International Conference on Fuzzy Systems & Knowledge Discovery*, July 23–25, Shenyang, China. pp 116–124.
- Yue, G., Barnes, S.J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, 59, 467–483.
- Zhang, Y.F., Li, H., Peng, L.H., & Hou, L.T. (2017). The usefulness classification algorithm and application of online reviews based on emotional semantic feature extraction. *Data Analysis and Knowledge Discovery*, 1(12), 74–83.
- Zhang, W.X., Wei, L., & Qi, J.J. (2005). The theory and method of attribute reduction of concept lattice. *Chinese Science E Series: Information Science*, 6, 628–639.
- Zhao, Y., Qin, B., Liu, T., & Tang, D. (2016). Social sentiment sensor: A visualization system for topic detection and topic sentiment analysis on microblog. *Multimedia Tools and Applications*, 75(15), 8843–8860.
- Zhi, X. (2002). Realization and optimization of association rule mining algorithm. *Computer Engineering & Applications*, 6(4), 341–357.
- Zheng, Q., Lu, Z., Yang, W., Zhang, M., Feng, Q., & Chen, W. (2013). A robust medical image segmentation method using KL distance and local neighborhood information. *Computers in Biology & Medicine*, 43(5), 459–470.
- Zhong, J., Zhang, S.F., Guo, W.L., & Li, X. (2018). TFLA: A Quality Analysis framework for student Generated contents. *Acta Electronica Sinica*, 46(9), 2201–2206.

This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).