Expert Systems With Applications 144 (2020) 113073

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Drawing openness to experience from user generated contents: An interpretable data-driven topic modeling approach

Yishi Zhang[a], Haiying Wei[b,*], Yaxuan Ran[c], Yang Deng[d], Dan Liu[e,f]

[a] School of Management, Wuhan University of Technology, Wuhan 430070, China
[b] School of Management, Jinan University, Guangzhou 510632, China
[c] School of Business Administration, Zhongnan University of Economics and Law, Wuhan 430073, China
[d] School of Management, Huazhong University of Science and Technology, Wuhan 430074, China
[e] China Research Center for Emergency Management, Wuhan University of Technology, Wuhan 430070, China
[f] School of Safety Science and Emergency Management, Wuhan University of Technology, Wuhan 430070, China

ARTICLE INFO

Article history:
Received 21 January 2019
Revised 13 October 2019
Accepted 2 November 2019
Available online 4 November 2019

Keywords:
Openness to experience
Interpretability
Topic modeling
Maximum-A-Posteriori estimation
Data-driven

ABSTRACT

Openness to experience, one of the essential individual characteristics, is of great theoretical and practical value in psychological and behavioral domains. Although typical machine learning methods can be utilized to extract individuals' openness to experience from the large-scale textual data like the unprecedented massive user generated contents (UGCs), they are often regarded as "black boxes" because they are unable to provide knowledge about the influential factors of openness to experience. This is of no help for us to investigate why a particular level of openness to experience is predicted for an individual. In addition, high dimensionality and sparseness of textual data impairs the performance of the typical machine learning method in extracting individuals' characteristics. In this study, we propose an interpretable data-driven mixture method for qualified modeling and predicting individuals' openness to experience. The proposed method extends the latent Dirichlet allocation (LDA) to overcome the problem of high dimensionality and sparseness in modeling the textual data, and can effectively extract two influential variables, namely, the topic preference and the expressed emotional intensity, to make an accurate prediction and to help us fully understand individuals' openness to experience lurking in the textual data. Experimental results indicate the effectiveness of the proposed method in drawing individuals' openness to experience, and also validate the predictive ability of topic preference and expressed emotional intensity which are indicated in psychological literature to be influential factors of openness to experience.

1. Introduction

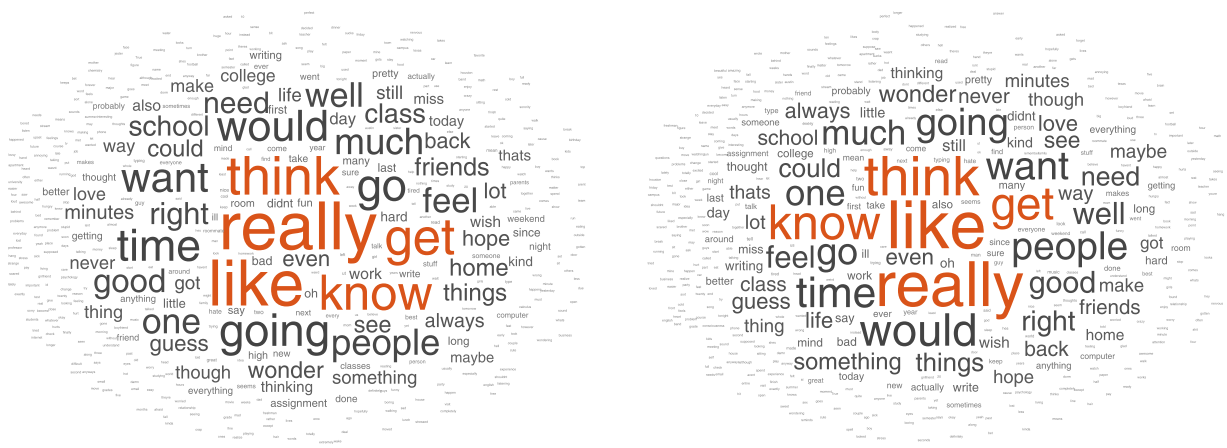
The fast-growing techniques for mobile networks and Internet-of-Things (IoT) provide the opportunity for users to act and express online naturally as in their real lives. The written language in the form of unstructured user generated contents (UGCs) accumulating online raises the availability of shaping the "digital profiles" of the users. Such massive "profiles" constitute an unprecedented source of rich information of the unique individual characteristics which are essential incentives for users' cognition and behaviors. To date, great efforts have been made to obtain insights on, for example, adoption behavior (Correa, Hinsley, & de Zúñiga, 2010), decision making (Weller, Ceschi, Hirsch, Sartori, & Geisler, 2018), and consumer preference (Büschken & Allenby, 2016; Ma, Chen, & Wei, 2017), from the individual characteristics lurking in UGCs.

1.1. Background

Openness to experience, one of the domains used to describe human personality in the Five Factor model, is a complex trait that reflects the psychological facet including fantasy, aesthetic sensitivity, attentiveness to inner feelings, preference for variety, and intellectual curiosity (McCrae & John, 1992). Psychological studies indicate that openness to experience is associated with emotions and sensations (Subramanian et al., 2018), and even subjective well-being (Steel, Schmidt, & Shultz, 2008). People high in openness are motivated to engage with perceptual and semantic information and are more likely to make novel associations

* Corresponding author.
E-mail addresses: easezh@126.com (Y. Zhang), tweihy@163.com (H. Wei), dengyang@hust.edu.cn (Y. Deng), liudan8575@whut.edu.cn (D. Liu).

https://doi.org/10.1016/j.eswa.2019.113073
0957-4174/



(a) Word cloud for $o = 0$ (b) Word cloud for $o = 1$

Fig. 1. An example of word clouds for essay dataset ($N = 2467$), where (a) shows the word for the sub-corpus in which the passages correspond to low level of openness to experience ($o = 0$), and (b) shows the word for the sub-corpus in which the passages correspond to high level of openness to experience ($o = 1$).

between remotely connected ideas (Christensen, Cotter, & Silvia, 2018; DeYoung, Grazioplene, & Peterson, 2012), while people low in openness are more comfortable with familiar and traditional experiences (Woo, Saef, & Parrigon, 2015). Openness to experience is an essential intrinsic characteristic that predicts preference heterogeneity and is one of the antecedent causes for decision making (Baer & Oldham, 2006; Islam, Rahman, & Hollebeek, 2017; Mulyanegara, Tsarenko, & Anderson, 2009; Rudd, Hildebrand, & Vhos, 2018; Zabkar, Arslanagic-Kalajdzic, Diamantopoulos, & Florack, 2017). Since within most of the UGCs lies knowledge of individuals' openness to experience that may be of great value to researchers, consumers, and marketers, it is imperative to develop interpretable text analysis techniques of data-driven discovery that can efficiently transform textual UGCs into structured data appropriate to real-world applications and paradigms of theoretical and empirical analysis (Humphreys & Wang, 2018).

Extant research mainly applies closed-vocabulary (i.e. dictionary-based) and open-vocabulary approaches for that purpose (Park et al., 2015). For the closed-vocabulary approach (e.g. Sumner, Byers, Boochever, & Park, 2012), the dynamics of UGCs result in frequent updating the associated dictionaries in order to keep performance. Additionally, the closed-vocabulary approach cannot capture the diversity of expression styles that are strongly associated with individual characteristics (Liu, Wang, Jiang, Sun, & Shang, 2018). By contrast, the open-vocabulary approach (e.g. Schwartz et al., 2013) employs techniques from computational linguistics and machine learning to extract a comprehensive collection of language features from the corpus being analyzed without predefined dictionaries (Park et al., 2015). To date, researchers have empirically validated that compared with the closed-vocabulary approach, such a data-driven mechanism of the open-vocabulary approach leads to more reliable extraction of individual characteristics (Humphreys & Wang, 2018; Iacobelli, Gill, Nowson, & Oberlander, 2011; Park et al., 2015; Schwartz et al., 2013).

Intuitively, individual characteristics are dependent on the topic preferences and the expression styles of the users. However, both two above-mentioned approaches rely mostly on word counts or frequencies within the corpus. Other potential informative variables, like aspect preferences and word co-occurrence with rich linguistic cues about individual characteristics, is seldom considered for analyzing and predicting individual characteristics. The problem with the simple analysis of word frequencies is that, it does not identify the combinations of predictive variables for which unique themes and points of differentiation are present. We

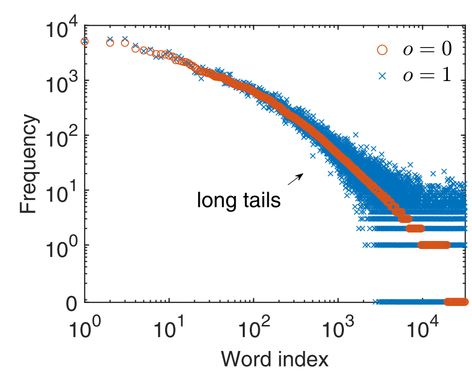


Fig. 2. The distributions of word frequencies for essay dataset ($N = 2467$). Word indices are assigned according to the descending order of the frequencies of the words in the sub-corpus in which all the passages are labeled as $o = 0$. Both axes are shown on a log-scale. The power-law-like relationship between word frequency and word index is evident from the near linear trend (in log-space) of this relationship.

take the corpus from the stream-of-consciousness essay dataset[1] as an example to elucidate the ineffectiveness of word frequencies, as shown in Figs. 1 and 2. The corpus contains 2467 passages (after preprocessing) written by 2467 individuals, where each individual has a specific level (low or high) of openness to experience. We use $o = 0$ ($o = 1$) to denote the low (high) level of openness to experience. Thus, each passage in the essay dataset alternatively corresponds to $o = 0$ or $o = 1$. Figs. 1(a) and 1(b) give the word clouds for the corpora within which all of the passages are written by the users with low and high levels of openness to experience, respectively. We can see that, the top-frequent words in Fig. 1(a) are nearly the same as those in Fig. 1(b), indicating that word frequency here is a very weak predictive variable. Fig. 2 shows the distributions of the word frequencies, where word indices are assigned according to the descending order of the frequencies of the words in the sub-corpus in which all the passages are labeled as $o = 0$. The similar power-law-like relationships for both $o = 0$ and $o = 1$ further manifest a poor predictive power of word frequency when modeling on these two similar distributions. It is important to note that, the oscillation of word frequencies with $o = 1$ existing in the long-tail of the distribution makes it different

[1] https://github.com/SenticNet/personality-detection.

from that with $o = 0$, implying that potential knowledge about the predictive power lurks in the long-tail of the distributions. Therefore, linguistic cues in the long-tail of the distributions need to be paid significant attention for improving predictive performance.

The long-tail of word frequency caused by high-dimensionality and sparseness of the words in the corpus serves as the main difficulty in text analysis for the traditional machine learning methods, e.g., support vector machines (SVM) (Rubin, Chambers, Smyth, & Steyvers, 2012). Being aware of this, the deep learning method (e.g., convolutional neural networks (CNN, Zeiler & Fergus, 2014) and recurrent neural networks RNN, Gers, Schraudolph, & Schmidhuber, 2003) and the topic model (e.g., the latent Dirichlet allocation LDA, Blei, Ng, & Jordan, 2003 and its variants Mcauliffe & Blei, 2008; Zhu, Ahmed, & Xing, 2012) are proposed to model word interactions. As a burgeoning approach, deep learning carries the torch for text mining and natural language processing (e.g., Araque, Corcuera-Platas, Snchez-Rada, & Iglesias, 2017; Do, Prasad, Maag, & Alsadoon, 2019), owing to its efficiency in feature extraction and representation (e.g., word embedding) for large volume of high dimensionality and sparse data. However, deep learning is referred as a typical “black box” approach (Lipton, 2018) which is unable to provide knowledge about the influential factors (features) of the dependent variable (i.e., the class) and thus of no help for us to investigate why a particular class label is predicted for an instance. The inability of researchers and practitioners to understand the deep learning model seems problematic and thus it is still not fully utilized in social science studies. By contrast, the topic model shows attractive potential for handling textual data in social science studies due to its interpretability. The automatically extracted topic preferences and other possible latent variables are of great help for the researchers and practitioners to further investigate individuals’ behaviors and characteristics (Humphreys & Wang, 2018). To date, although the topic model has been attracting increasing attentions in psychological and behavioral studies (e.g., Hasan, Ferguson, & Koning, 2015; Humphreys & Wang, 2018; Liu, Wang, & Jiang, 2016; Park et al., 2015; Shi, Lee, & Whinston, 2016; Wang & Chaudhry, 2018), it is applied only in a few of them (Liu et al., 2016; Park et al., 2015) for mining the individual characteristics or the personality traits. In addition, other potential predictive features, like the expressed emotion lurking in the text which is relevant to openness to experience (Subramanian et al., 2018; Tok, Koyuncu, Dural, & Catikkas, 2010), are still neglected in individual characteristics prediction.

1.2. Contributions and outline

Different from extant studies, we focus on modeling and predicting individuals’ openness to experience using two influential factors, namely, the topic preference and the expressed emotional intensity, lurking in the UGCs. Specifically, we propose in this study a Bayesian topic-emotion-openness mixture model to effectively predict the trait of openness to experience, and apply the Gibbs sampler and the Maximum-A-Posteriori (MAP) estimation for model inference and prediction in a data-driven manner. The proposed method has the following advantages. Firstly, the study extends LDA to extract both topic preference and expressed emotional intensity for openness modeling, where LDA-based framework is efficient to be adapted to handle large volume and sparse data. Secondly, this study proposes an interpretable data-driven method for qualified modeling and predicting individuals’ openness to experience, which strengthens the links between empirical psychological research and machine learning techniques. Thirdly, LDA-based framework utilizes an unsupervised Bayesian learning approach to capture context specific topics and emotion cues, where the interaction of words that may lead to opposite meanings in contrast with the individual words can be taken into account. Fourthly, the proposed method does not need any prior information about the structure of the text or grammatical properties of the language. The main contributions of this paper to literature is listed as follows.

(1) To the best of our knowledge, it is the first attempt to concentrate on predicting individuals’ openness to experience with the interpretable topic-emotion-openness mixture model without the limitation of the closed-vocabulary approach.
(2) The Maximum-A-Posteriori (MAP) estimation is novelly embedded into a three-level Bayesian generation process in our method for modeling and predicting openness to experience in a data-driven manner.
(3) The proposed method extracts two influential factors, namely, the topic preference and the expressed emotional intensity, for a reasonable understanding of openness to experience lurking in the textual data.

The remainder of the paper is organized as follows: Section 2 briefly surveys the related work on personality trait prediction and topic modeling. Section 3 briefly describe LDA. Then Section 4 proposes a topic-emotion-openness mixture model and its Bayesian inference which can effectively discover the knowledge of openness to experience lurking in the textual data, and then Section 5 introduces an approximate inference for the proposed model. In Section 6, experimental results are given to evaluate the effectiveness of proposed method comparing with the existing methods, discussions are then presented. Finally, Section 7 gives the conclusions and the future research directions.

2. Related work

As mentioned previously, the existing methods mainly apply closed-vocabulary or open-vocabulary approaches to deal with textual data for individual characteristics analysis and prediction. The closed-vocabulary approach, as its name implies, counts and evaluates the individual’s use of words via predefined categories of words (i.e. the dictionary) (Schwartz et al., 2013). A representative implementation of this approach, called the Linguistic Inquiry and Word Count (LIWC, Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007), automatically counts word frequencies within the corpus for multiple predefined categories associated with psychological, behavioral, and linguistical aspects. Within recent psychological and behavioral studies, the closed vocabulary approach like LIWC has become particularly popular because of its ease of use and helpfulness to interpretability. Pennebaker and King (1999) apply a closed-vocabulary approach to analyze the textual data collected from their experiments, and conclude that language use can reflect individual characteristics. Golbeck, Robles, and Turner (2011b) use LIWC to analyze the textual data from the personal profiles and messages of Facebook users, and indicate that the users’ characteristics can be predicted from the public information they share online. Similar predictive models of individual characteristics have also built with the closed-vocabulary approach from Twitter or Facebook (Golbeck, Robles, Edmondson, & Turner, 2011a; Holtgraves, 2011; Mehl, Gosling, & Pennebaker, 2006; Sumner et al., 2012). For example, Mehl et al. (2006) investigate the relationship between users’ topic preference and their openness to experience from a closed-vocabulary based study, and conclude that people low in openness prefer more topics about social process. Their subtle and meaningful findings shed light upon the prediction of individuals’ openness to experience from UGCs.

By contrast, the open-vocabulary method does not depend on pre-defined dictionaries and can automatically generate more numerous and richer features rather than word frequency from the corpus. It is originally developed from the field of computational linguistics and machine learning and has been indicated to be more effective for analysis and prediction of individual characteristics from online textual data (Alam, Stepanov, & Riccardi, 2013; Iacobelli et al., 2011; Liu et al., 2016; Schwartz et al., 2013). For example, Iacobelli et al. (2011) apply Support Vector Machines (SVM) to predict individual characteristics using massive textual data from blogs, and indicate that the open-vocabulary approach significantly outperforms the closed-vocabulary approach within the online corpora. Similar results are also obtained by a series of works that use machine learning based open-vocabulary approaches, including decision trees (Wald, Khoshgoftaar, & Sumner, 2012), naïve Bayesian classifier (Alam et al., 2013), SVM (Farnadi, Zoghbi, Moens, & Cock, 2013; Tandera, Hendro, Suhartono, Wongso, & Prasetio, 2017), and topic models (Liu et al., 2016; Park et al., 2015; Schwartz et al., 2013).

Topic as a semantic feature of the textual data is first applied as a predictive variable of individual characteristics by Schwartz et al. (2013), who apply the representative topic model LDA (Blei et al., 2003) and Pairwise Mutual Information (PMI) to predict individual characteristics using massive textual data from Facebook. The prediction results indicate that topics extracted by LDA significant outperform the categories of LIWC in predicting individual characteristics from online textual data. Park et al. (2015) apply LDA to extract topics from the textual data obtaining through the myPersonality Facebook application[2] for personality assessment, indicating that the textual data in social media can be harnessed to create a valid and reliable measure of personality. Liu et al. (2016) propose a supervised learning method slightly different from LDA, arguing that each concrete topic assigned to a document randomly generates not only a word but also personality traits. Their experimental results show feasibility of their method as an alternative means for predicting individual characteristics.

The topic models have prevailed to date since LDA was first proposed by Blei et al. (2003). To date, many variants of LDA are proposed for various tasks in different scenarios, with not only the corpus but also the additional information as either the informative priori or the observed variables, including timestamps (Wang & McCallum, 2006), authors (Rosen-Zvi, Chemudugunta, Griffiths, Smyth, & Steyvers, 2010), and citations (Nallapati, Ahmed, Xing, & Cohen, 2008) among others. Other variants are developed by embedding new latent variables like sentiment (Jo & Oh, 2011; Lin & He, 2009) and objectivity (Wang et al., 2018), or by adding constraints that are fit for the corpus from specific scenarios, e.g. sentence-based topic generation constraint (Büschken & Allenby, 2016; Jo & Oh, 2011). For learning categories, LDA is also modified to multiple supervised variants, e.g. supervised-LDA (Mcauliffe & Blei, 2008) and labeled-LDA (Ramage, Hall, Nallapati, & Manning, 2009). Compared with the conventional machine learning methods only considering word frequencies or LIWC categories, the topic models generate topics of which the distribution can reflect the aspect preference of the users that contains rich information of individual characteristics (Mehl et al., 2006). However, as introduced above, limited attentions are received to apply or extend the topic models to extract topics and topic-related predictive variables from textual data for individual characteristics analysis and prediction.

3. Preliminaries

In this section, we briefly introduce LDA as the preliminaries to the proposed method. LDA is an unsupervised Bayesian generative model that can extract a fixed number of latent topics that appear across multiple documents (i.e. samples in the corpus) without knowledge of pre-labeled topics. The key assumption of LDA is the conditional independence between documents and words. That is, documents are dependent only on topics while topics are dependent only on words. Each document consists of its own mixture of topics, and each topic is characterized by a multinomial distribution over words. Suppose that the corpus contains V different words, D documents, and T topics, where the d-th document contains N_d words, the generative process of LDA can be described as follows. The multinomial distribution over topics for the d-th document, $\boldsymbol{\theta}_d \in \mathbb{R}^{T\times 1}$, and the multinomial distribution over words for the t-th topic, $\boldsymbol{\phi}_t \in \mathbb{R}^{V\times 1}$, are drawn from two Dirichlet distributions denoted as $\mathrm{Dir}(\boldsymbol{\alpha})$ and $\mathrm{Dir}(\boldsymbol{\beta})$, respectively, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the parameters of the Dirichlet distributions. The concrete topics $\mathbf{z}_d = \{z_{d,n}\}_{n=1}^{N_d}$ associated with the words in the d-th document are repeatedly sampled from $\boldsymbol{\theta}_d$. Then the concrete words $\mathbf{w}_d = \{w_{d,n}\}_{n=1}^{N_d}$ are repeatedly sampled from $\boldsymbol{\phi}_{z_{d,n}}$. Formally, the above generative process is shown in Algorithm 1 . Table 1 gives the notations of the parameters and the variables used in LDA.

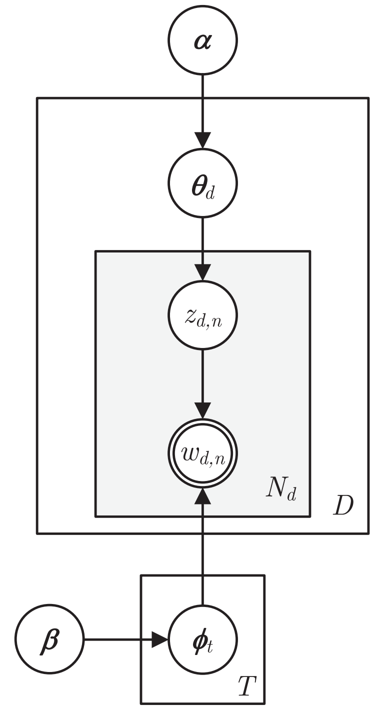


Fig. 3. Graphical diagram with plate notations of LDA.

The corresponding graphical presentation with plate notations of LDA is shown in Fig. 3, where the plates (i.e. rectangles) represent replicates with respect to document C_d $(d = 1, \ldots, D)$, word w_{dn} $(n = 1, \ldots, N_d)$, and topic z_t $(t = 1, \ldots, T)$, respectively. Note that LDA does not impose any relationships of variables related to the plates; It assumes that the concrete topics $\mathbf{z}$ are independent with each other.

LDA is a typical *bags-of-words* model, where a document is represented as the multiset of its words, keeping only multiplicity (word frequency) without information of grammar and word order. The presence of the latent topics in the document are in essence determined by the frequencies and interactions of the words. These topics introduce a set of word interactions into an analysis so that words with high topic probabilities are jointly predicted to be present. The objects of inference of LDA are the parameter matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ that contain all the information of the probabilities of the topics for each document d and the corresponding words

[2] http://www.mypersonality.org/wiki.

Algorithm 1: The generative process of LDA.

/* Generate posterior parameters $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_t$ */

1. For each document $Document_d$ $(1 \le d \le D)$:
   1.1 Draw a multinomial distribution: $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha}) \quad \forall d$ i.i.d.
2. For each topic $Topic_t$ $(1 \le t \le T)$:
   2.1 Draw a multinomial distribution: $\boldsymbol{\phi}_t \sim \mathrm{Dir}(\boldsymbol{\beta}) \quad \forall t$ i.i.d.

/* Generate the corpus */

1. For each document $Document_d$ $(1 \le d \le D)$:
   1.1 For each item $Item_{d,n}$ $(1 \le n \le N_d)$:
      1.1.1 Draw a topic z_{dn}: $z_{dn} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d) \quad \forall d, n$ i.i.d
      1.1.2 Draw a word w_{dn}: $w_{dn} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{dn}}) \quad \forall d, n$ i.i.d

Table 1
Parameters and variables used in LDA.

Notation	Description
D	Number of documents
T	Number of topics
N_d	Number of words in the d-th document, $d = 1, \ldots, D$
V	Number of words in the vocabulary
$\boldsymbol{\alpha}$	Dirichlet prior parameters for $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D)$
$\boldsymbol{\beta}$	Dirichlet prior parameters for $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_T)$
$\boldsymbol{\theta}_d$	Multinomial distribution over topics for the d-th document, $d = 1, \ldots, D$
$\boldsymbol{\phi}_t$	Multinomial distribution over words for the t-th topic, $t = 1, \ldots, T$
$z_{d,n}$	Topic of the n-th word in the d-th document, $n = 1, \ldots, N_d$, $d = 1, \ldots, D$
$w_{d,\,n}$	The n-th word in the d-th document, $n = 1, \ldots, N_d$, $d = 1, \ldots, D$

for each topic t. A model involving T topics is associated with a $T \times D$ probability matrix $\boldsymbol{\Theta}$ and a $V \times T$ probability matrix $\boldsymbol{\Phi}$.

4. The proposed model

LDA provides an effective way to model interactions of words. And this inspires researchers to extract topic and topic-like variable using the LDA-based framework that may be of great help in individual characteristic analysis. In this section, we extends LDA to a topic-emotion-openness mixture model for extracting individuals' openness to experience from their words.

Intuitively, the expressed emotional intensity from an individual's words is a topic-like variable that is strongly related to the combinations of words. It is thus possible to be extracted by observed word co-occurrences using the LDA-based framework. However, as a complex individual characteristics, openness to experience seems not able to be effectively reflected by low-level interactions of words, i.e. word co-occurrences, but is more likely to be reflected by the features relevant to high-level interactions of words, notably topic preference (Mehl et al., 2006) and expressed emotion (Subramanian et al., 2018; Tok et al., 2010). Being aware of this, our proposed method holds assumptions that openness to experience (a) is dependent on topic preference and emotional intensity extracted from words, and (b) is conditional independent to words given the distributions of latent topic and emotional intensity. A graphical model of the proposed method is represented in Fig. 4 with the notations defined in Table 2.

The proposed method extend LDA by adding an additional emotional intensity layer between the document and the topic layer, and an additional openness to experience layer between the topic distribution layer and the emotional intensity layer. Hence, the proposed model is effectively a five-layer Bayesian model, where emotional intensity labels are associated with openness to experience labels and documents, under which (a) topics are associated with both openness to experience labels and emotional intensity labels and (b) words are associated with both emotional intensity labels and topics. The formal definition of the generative process corresponding to the proposed model is represented in Algorithm 2 .

It is worth noting that, the topic-emotion-document distribution in the proposed method, i.e. $\boldsymbol{\theta}$, is different from that in LDA, because in our method each document is associated with S topic-document distributions, where each topic-document distribution corresponds to a concrete emotional intensity label with the same number of topics. Similarly, the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are also different from those in LDA, because each $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ are associated respectively with K hyperparameters, where each $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ correspond respectively to a concrete openness to experience label with the same number of hyperparameters of the topics. These characteristics in essence provide means for our method to extract openness to experience from the distributions of topics and emotional intensity.

According to the graphical model represented in Fig. 4, the joint distribution of all variables for the corpus can be formulated as follows:

$$\begin{aligned} &p(\mathbf{w}, \mathbf{z}, \mathbf{v}, \mathbf{o}, \boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\Pi} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}) \\ &\quad = p(\mathbf{w}|\boldsymbol{\Phi}, \mathbf{v}, \mathbf{z}) \cdot p(\boldsymbol{\Phi}|\boldsymbol{\beta}) \cdot p(\mathbf{z}|\mathbf{v}, \boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta}|\mathbf{o}, \boldsymbol{\alpha}) \cdot p(\mathbf{v}|\boldsymbol{\Pi}) \\ &\quad\quad \cdot p(\boldsymbol{\Pi}|\mathbf{o}, \boldsymbol{\gamma}) \cdot p(\mathbf{o}|\boldsymbol{\lambda}) \cdot p(\boldsymbol{\Lambda}|\boldsymbol{\xi}) \end{aligned} \tag{1}$$

where $p(\boldsymbol{\Phi}|\boldsymbol{\beta})$ corresponds to the generation process represented in the item 1.1 in Algorithm 2; $p(\boldsymbol{\Pi}|\mathbf{o}, \boldsymbol{\gamma})$, $p(\boldsymbol{\Theta}|\mathbf{o}, \boldsymbol{\alpha})$, and $p(\boldsymbol{\Lambda}|\boldsymbol{\xi})$ correspond to the generation processes represented in items 2.2, 2.3, and 2.4 in Algorithm 2, respectively; $p(\boldsymbol{o}|\boldsymbol{\lambda})$, $p(\mathbf{v}|\boldsymbol{\Pi})$, $p(\mathbf{z}|\mathbf{v}, \boldsymbol{\Theta})$, and $p(\mathbf{w}|\boldsymbol{\Phi}, \mathbf{v}, \mathbf{z})$ correspond to the generation processes represented in items 3.1.1, 3.1.2, and 3.1.3 in Algorithm 2, respectively. In the next section, we will use the collapsed Gibbs sampler and the Maximum-A-Posteriori (MAP) estimation for model inference and prediction.

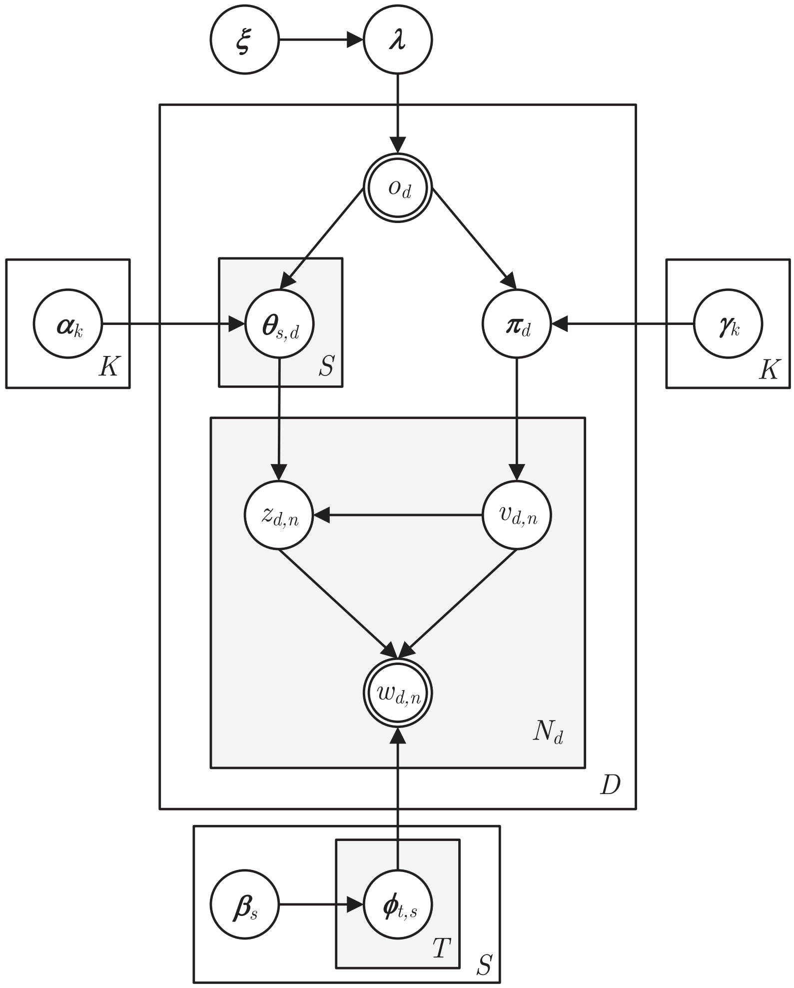


Fig. 4. Graphical diagram with plate notations of the proposed model.

Table 2
Parameters and variables used in the proposed method.

Notation	Description
D	Number of documents
T	Number of topics
N_d	Number of words in the d-th document, $d = 1, \ldots, D$
S	Number of levels (labels) of emotional intensity
K	Number of levels (labels) of openness to experience
V	Number of words in the vocabulary
$\boldsymbol{\alpha}_k$	Dirichlet prior parameters for $\boldsymbol{\Theta}_k = \left(\boldsymbol{\theta}_{1,n_1^k}, \ldots, \boldsymbol{\theta}_{S,n_{D_k}^k}\right)$, where D_k denotes the number of the documents labeled with the k-th level of openness to experience and n_j^k denotes index of such the j-th document, $k = 1, \ldots, K$, $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K)$
$\boldsymbol{\beta}_s$	Dirichlet prior parameters for $\boldsymbol{\Phi}_s = \left(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{T,s}\right)$, $s = 1, \ldots, S$, $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_S)$
$\boldsymbol{\gamma}_k$	Dirichlet prior parameters for $\boldsymbol{\Pi}_k = \left(\boldsymbol{\pi}_{n_1^k}, \ldots, \boldsymbol{\pi}_{n_{D_k}^k}\right)$, where D_k denotes the number of the documents labeled with the k-th level of openness to experience and n_j^k denotes index of such the j-th document, $k = 1, \ldots, K$, $\boldsymbol{\Pi} = (\boldsymbol{\Pi}_1, \ldots, \boldsymbol{\Pi}_K)$
$\boldsymbol{\xi}$	Dirichlet prior parameters for $\boldsymbol{\Lambda} = \left(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_D\right)$
$\boldsymbol{\lambda}_d$	Multinomial distribution over documents, $d = 1, \ldots, D$
$\boldsymbol{\theta}_{s,d}$	Multinomial distribution over topics for the s-th level of emotional intensity and the d-th document, $s = 1, \ldots, S$, $d = 1, \ldots, D$
$\boldsymbol{\pi}_d$	Multinomial distribution over levels of the emotional intensity for the d-th document, $d = 1, \ldots, D$
$\boldsymbol{\phi}_{t,s}$	Multinomial distribution over words for the t-th topic and the s-th level of emotional intensity, $t = 1, \ldots, T$, $s = 1, \ldots, S$
o_d	Openness to experience for the d-th document, $d = 1, \ldots, D$
$z_{d,n}$	Topic of the n-th word in the d-th document, $n = 1, \ldots, N_d$, $d = 1, \ldots, D$
$w_{d,\,n}$	The n-th word in the d-th document, $n = 1, \ldots, N_d$, $d = 1, \ldots, D$
$v_{d,\,n}$	Level of emotional intensity of the n-th word in the d-th document, $n = 1, \ldots, N_d$, $d = 1, \ldots, D$

Algorithm 2: The generative process of the proposed model.

1. For each topic $Topic_t$ $(1 \le t \le T)$:
   1.1 For each level of emotional intensity $Emotion_s$ $(1 \le s \le S)$:
      1.1.1 Draw a multinomial distribution: $\boldsymbol{\phi}_{t,s} \sim \mathrm{Dir}(\boldsymbol{\beta}_s)$ $\forall t, s$ i.i.d.
2. For each document $Document_d$ $(1 \le d \le D)$:
   2.1 Draw a level of openness to experience: $o_d \sim \mathrm{Multinomial}(\boldsymbol{\lambda}_d)$ $\forall d$ i.i.d
   2.2 Draw a multinomial distribution: $\boldsymbol{\pi}_d \sim \mathrm{Dir}(\boldsymbol{\gamma}_{o_d})$ $\forall d$ i.i.d.
   2.3 For each level of emotional intensity $Emotion_s$ $(1 \le s \le S)$:
      2.3.1 Draw a multinomial distribution: $\boldsymbol{\theta}_{s,d} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{o_d})$ $\forall s, d$ i.i.d.
   2.4 Draw a multinomial distribution: $\boldsymbol{\lambda}_d \sim \mathrm{Dir}(\boldsymbol{\xi})$ $\forall d$ i.i.d
3. For each document $Document_d$ $(1 \le d \le D)$:
   3.1 For each item $Item_{d,n}$ $(1 \le n \le N_d)$:
      3.1.1 Draw a level of emotional intensity: $v_{d,n} \sim \mathrm{Multinomial}(\boldsymbol{\pi}_d)$ $\forall d, n$ i.i.d
      3.1.2 Draw a topic: $z_{d,n} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_{v_{d,n},d})$ $\forall d, n$ i.i.d
      3.1.3 Draw a word: $w_{d,n} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{d,n},v_{d,n}})$ $\forall d, n$ i.i.d

5. Inference and prediction

In the proposed method, there are four sets of latent variables that we need to infer from word assignments observed in the corpus, including the joint topic-emotional intensity-document distributions $\boldsymbol{\theta}_{s,d}$ (hereafter called topic-emotion-document distributions), the joint emotional intensity-topic-word distributions $\boldsymbol{\phi}_{t,s}$ (hereafter called emotion-topic-word distributions), the emotional intensity-document distributions $\boldsymbol{\pi}_d$ (hereafter called emotion-document distributions), and the openness to experience-document distributions $\boldsymbol{\lambda}_d$ (hereafter called openness-document distributions). Suppose we have obtained the final assignments of **z, v** for all words and the final assignment of **o** for all documents in the corpus, the four distributions $\boldsymbol{\theta}_{s,d}$, $\boldsymbol{\phi}_{t,s}$, $\boldsymbol{\pi}_d$, and $\boldsymbol{\lambda}_d$ can thus be estimated in terms of all observed (w, z, v)-pairs and (o, d)-pairs. To this end, $\boldsymbol{\theta}_{s,d}$, $\boldsymbol{\pi}_d$, $\boldsymbol{\phi}_{t,s}$, and $\boldsymbol{\lambda}_d$ should be first integrated out from Eq. (1) and then the Gibbs sampling can be applied to generate (w, z, v)-pairs and (o, d)-pairs for the estimate of $\boldsymbol{\theta}_{s,d}$, $\boldsymbol{\pi}_d$, $\boldsymbol{\phi}_{t,s}$, and $\boldsymbol{\lambda}_d$ after convergence. We focus on the following formulation with Eq. (1)

$$\begin{aligned} p(\mathbf{w}, \mathbf{z}, \mathbf{v}, \mathbf{o}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}) &= \iiint_{\boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\Pi}, \boldsymbol{\Lambda}} p(\mathbf{w}, \mathbf{z}, \mathbf{v}, \mathbf{o}, \boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\Pi}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}) \cdot \mathrm{d}\boldsymbol{\Theta}\mathrm{d}\boldsymbol{\Phi}\mathrm{d}\boldsymbol{\Pi}\mathrm{d}\boldsymbol{\Lambda} \\ &= p(\mathbf{w}|\mathbf{z}, \mathbf{v}, \boldsymbol{\beta}) \cdot \underbrace{\left(\int_{\boldsymbol{\Theta}} p(\mathbf{z}|\mathbf{v}, \boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta}|\mathbf{o}, \boldsymbol{\alpha}) \cdot \mathrm{d}\boldsymbol{\Theta}\right)}_{\mathrm{I}} \\ &\quad \cdot \underbrace{\left(\int_{\boldsymbol{\Pi}} p(\mathbf{v}|\boldsymbol{\Pi}) \cdot p(\boldsymbol{\Pi}|\mathbf{o}, \boldsymbol{\gamma}) \cdot \mathrm{d}\boldsymbol{\Pi}\right)}_{\mathrm{II}} \cdot p(\mathbf{o}|\boldsymbol{\xi}), \end{aligned} \tag{2}$$

where

$$p(\mathbf{w}|\mathbf{z}, \mathbf{v}, \boldsymbol{\beta}) = \prod_{t=1}^{T}\prod_{s=1}^{S} \frac{\sum_{i=1}^{V}\Gamma(n_{i,t,s} + \beta_{i,t,s})}{\Gamma(n_{ts} + \sum_{i=1}^{V}\beta_{i,t,s})} \cdot \frac{\Gamma(\sum_{i=1}^{V}\beta_{i,t,s})}{\prod_{i=1}^{V}\Gamma(\beta_{i,t,s})}, \tag{3}$$

herein $n_{w,\,z,\,v}$ and $n_{z,\,v}$ denote the counts of (w, z, v)-pairs and (z, v)-pairs in the corpus, respectively. The derivation of Eq. (3) is similar to LDA and thus is omitted. For I and II in Eq. (2), latent parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Pi}$ are not conjugately distributed with respect to the topic **z** and the emotional intensity **v** because of the existence of the conditioning variable **o**. The integrations are thus intractable since it is hard to model the relationship between **o** and $\boldsymbol{\Theta}$ (and $\boldsymbol{\Pi}$). Although this dilemma can be avoid by assuming a certain relationship between $\boldsymbol{\alpha}$ and **o** and the one between $\boldsymbol{\gamma}$ and **o**, formulated as $f(\boldsymbol{\alpha})$ and $f(\boldsymbol{\gamma})$, respectively, for replacing **o** in the integration with $f(\boldsymbol{\alpha})$ and $f(\boldsymbol{\gamma})$ similarly to what labeled-LDA has done (for more details, see Ramage et al., 2009), the performance of the model may be restricted if there is no accurate prior knowledge for that assumption. We thus take the assumption of conditional independence between $\boldsymbol{\alpha}/\boldsymbol{\gamma}$ and **o**, and predict **o** in a data-driven manner with the Maximum-A-Posteriori (MAP) estimation: When generating **z** and **v**, no prior relationships are assumed beforehand between $\boldsymbol{\alpha}/\boldsymbol{\gamma}$ and **o**, i.e. $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \ldots = \boldsymbol{\alpha}_K$ and $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2 = \ldots = \boldsymbol{\gamma}_K$. After that, the MAP is applied to predict **o** by learning the relationship between the estimated $\boldsymbol{\theta}/\boldsymbol{\pi}$ and **o**. This procedure is data-driven because the relationships are not predefined but automatically learned from the corpus.

5.1. Sampling z and v for each word w in the corpus

We first obtain $\boldsymbol{\Theta}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Pi}$ by Gibbs sampling. Denote the topic of the current word w_i in the current document is z_i and $\mathbf{w} = \{w_i, \mathbf{w}_{\neg i}\}$, $\mathbf{z} = \{z_i, \mathbf{z}_{\neg i}\}$, and $\mathbf{v} = \{v_i, \mathbf{v}_{\neg i}\}$. With Eq. (2) and the above-mentioned assumption, the conditional distribution of z_i is proportional to

$$p(z_i|\mathbf{z}_{\neg i}, \mathbf{w}, \mathbf{v}, \mathbf{o}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}) \propto \frac{p(\mathbf{w}|\mathbf{z}, \mathbf{v}, \boldsymbol{\beta})}{p(\mathbf{w}_{\neg i}|\mathbf{z}_{\neg i}, \mathbf{v}_{\neg i}, \boldsymbol{\beta})} \cdot \underbrace{\frac{\int_{\boldsymbol{\Theta}} p(\mathbf{z}|\mathbf{v}, \boldsymbol{\Theta}, \boldsymbol{\alpha}) \cdot p(\boldsymbol{\Theta}|\mathbf{o}) \cdot \mathrm{d}\boldsymbol{\Theta}}{\int_{\boldsymbol{\Theta}} p(\mathbf{z}_{\neg i}|\mathbf{v}_{\neg i}, \boldsymbol{\Theta}, \boldsymbol{\alpha}) \cdot p(\boldsymbol{\Theta}|\mathbf{o}) \cdot \mathrm{d}\boldsymbol{\Theta}}}_{\hbar(z_i)} \tag{4}$$

where

$$\begin{aligned} \hbar(z_i) &= \frac{\prod_{d=1}^{D}\prod_{j=1}^{N_d}\int_{\boldsymbol{\theta}_d} p(z_{dj}|v_{dj}, \boldsymbol{\theta}_d, \boldsymbol{\alpha}) \cdot p(\boldsymbol{\theta}_d|o_d) \cdot \mathrm{d}\boldsymbol{\theta}_d}{\prod_{d=1}^{D}\prod_{j=1, j\neq i}^{N_d}\int_{\boldsymbol{\theta}_d} p(z_{dj}|v_{dj}, \boldsymbol{\theta}_d, \boldsymbol{\alpha}) \cdot p(\boldsymbol{\theta}_d|o_d) \cdot \mathrm{d}\boldsymbol{\theta}_d} \\ &= \int_{\boldsymbol{\theta}_d} p(z_i|v_i, \boldsymbol{\theta}_d, \boldsymbol{\alpha}) \cdot p(\boldsymbol{\theta}_d|o_d) \cdot \mathrm{d}\boldsymbol{\theta}_d \end{aligned} \tag{5}$$

and $\boldsymbol{\theta}_d = (\boldsymbol{\theta}_{1,d}, \ldots, \boldsymbol{\theta}_{S,d})$. Since $p(z_i|v_i, \boldsymbol{\theta}_d, \boldsymbol{\alpha})$ is fully determined by the prior variables v_{dj} and $\boldsymbol{\theta}_d$, it can be straightforwardly estimated as $\frac{n_{z_i,v_i}^{d,\neg i}}{n_d}$ (hereinafter $n_{\cdot}^{\neg i}$ denote the count of the certain patterns in the corpus excluding the i-th item, $n_{z_i,v_i}^{d,\neg i}$ and $n_d^{\neg i}$ denote the counts of (z_i, v_i)-pairs and words in the d-th document without the topic, emotional intensity label, and word assignments for the i-th item, respectively). To fully take into account the stochasticity of $\boldsymbol{\theta}_d$, we may finally get the expression of $p(z_i|v_i, \boldsymbol{\theta}_d, \boldsymbol{\alpha})$ in terms of the property of conjugate priori of mutinomial distribution as

$\frac{n^{d,\neg i}_{z_i,v_i}+\alpha_{z_i,v_i}}{n^{d,\neg i}_{v_i}+\sum_{t=1}^{T}\alpha_{t,v_i}}$. Hence we have

$$\hbar(z_i) = p(z_{dj}|v_{dj},\boldsymbol{\theta}_d,\boldsymbol{\alpha})\cdot\int_{\boldsymbol{\theta}_d} p(\boldsymbol{\theta}_d|o_d)\cdot \mathrm{d}\boldsymbol{\theta}_d = \frac{n^{d,\neg i}_{z_i,v_i}+\alpha_{z_i,v_i}}{n^{d,\neg i}_{v_i}+\sum_{t=1}^{T}\alpha_{t,v_i}}. \tag{6}$$

With Eqs. (3) and (5),

$$p(z_i|\mathbf{z}_{\neg i},\mathbf{w},\mathbf{v},\mathbf{o},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\xi}) \propto \frac{(n^{\neg i}_{w_i,z_i,v_i}+\beta_{w_i,z_i,v_i})(n^{d,\neg i}_{z_i,v_i}+\alpha_{z_i,v_i})}{n^{\neg i}_{z_i,v_i}+\sum_{j=1}^{V}\beta_{j,z_i,v_i}}. \tag{7}$$

Eq. (7) will be used as the sampling function of z_i. Similarly, the sampling function of v_i has the following form

$$p(v_i|\mathbf{z},\mathbf{w},\mathbf{v}_{\neg i},\mathbf{o},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\xi}) \propto \frac{(n^{\neg i}_{w_i,z_i,v_i}+\beta_{w_i,z_i,v_i})(n^{d,\neg i}_{v_i}+\gamma_{v_i})}{n^{\neg i}_{z_i,v_i}+\sum_{j=1}^{V}\beta_{j,z_i,v_i}} \tag{8}$$

5.2. A data-driven procedure for openness to experience prediction

5.2.1. Inference on the testing corpus

Gibbs sampling on the training corpus will finally lead to stable $\phi_{w,\,z,\,v}$, $\theta_{z,\,v,\,d}$, and $\pi_{v,\,d}$ with the formula of

$$\phi_{w,z,v} = \frac{n_{w,z,v}+\beta_{w,z,v}}{n_{z,v}+\sum_{j=1}^{V}\beta_{j,z,v}}, \tag{9}$$

$$\theta_{z,v,d} = \frac{n^{d}_{z,v}+\alpha_{z,v}}{n_{v,d}+\sum_{t=1}^{T}\alpha_{t,v}}, \tag{10}$$

and

$$\pi_{v,d} = \frac{n^{d}_{v}+\gamma_{v}}{n_d+\sum_{s=1}^{S}\gamma_s}. \tag{11}$$

We now apply $\phi_{w,\,z,\,v}$ as the priori for the inference on the testing corpus. Specifically, given the d^{test}-th test document, the conditional distribution of the associated topic for the i-th word w'_i is formulated as

$$\begin{aligned} &p(z'_i|\mathbf{z}'_{\neg i},\mathbf{z},\mathbf{w},\mathbf{v},\mathbf{o},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\xi}) \\ &\propto \frac{(n^{\neg i}_{w'_i,z'_i,v'_i}+n^{d^{\text{test}},\neg i}_{w'_i,z'_i,v'_i}+\beta_{w'_i,z'_i,v'_i})(n^{d^{\text{test}}}_{z'_iv'_i}+\alpha_{z'_iv'_i})}{n^{\neg i}_{z'_i,v'_i}+n^{d^{\text{test}},\neg i}_{z'_i,v'_i}+\sum_{j=1}^{V}\beta_{j,z'_i,v'_i}} \\ &\doteq \frac{(n^{\neg i}_{w'_i,z'_i,v'_i}+\beta_{w'_i,z'_i,v'_i})(n^{d^{\text{test}},\neg i}_{z'_iv'_i}+\alpha_{z'_iv'_i})}{n^{\neg i}_{z'_i,v'_i}+\sum_{j=1}^{V}\beta_{j,z'_i,v'_i}} \\ &= \phi^{\text{train}}_{w'_i,z'_i,v'_i}\cdot(n^{d^{\text{test}},\neg i}_{z'_iv'_i}+\alpha_{z'_iv'_i}), \end{aligned} \tag{12}$$

where w'_i, z'_i, v'_i are the i-th word, the associated topic, and the associated emotional intensity in the d^{test}-th document, respectively. $n^{d^{\text{test}}}_{w'_i,z'_i,v'_i}$ and $n^{d^{\text{test}}}_{z'_i,v'_i}$ denote the counts of (w'_i,z'_i,v'_i)-pairs and (z'_i,v'_i)-pairs in the d^{test}-th document, respectively. The second approximation term of (12) holds because $n^{d^{\text{test}}}_{w'_i,z'_i,v'_i} \ll n^{w,z,v}_{w'_i,z'_i,v'_i}$ and $n^{d^{\text{test}}}_{z'_i,v'_i} \ll n^{z,v}_{z'_i,v'_i}$. We use such the approximation in the proposed method to obtain a faster sampling process for test samples. Similarly, the conditional distribution of the associated emotional intensity v'_i has the following form

$$p(v'_i|\mathbf{v}'_{\neg i},\mathbf{z},\mathbf{w},\mathbf{v},\mathbf{o},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\xi}) \mathrel{\dot\propto} \phi^{\text{train}}_{w'_i,z'_i,v'_i}\cdot(n^{d^{\text{test}},\neg i}_{v'_i}+\gamma_{v'_i}). \tag{13}$$

With the sampling functions being formulated, we can finally get $\boldsymbol{\Theta}^{\text{test}}$, $\boldsymbol{\Phi}^{\text{test}}$, and $\boldsymbol{\Pi}^{\text{test}}$ after sufficient iterations of Gibbs sampling.

5.2.2. Modeling and predicting openness to experience with MAP

Given $\boldsymbol{\theta}_d$, $\boldsymbol{\xi}_d$, $\boldsymbol{\alpha}$, $\boldsymbol{\xi}$, and $\boldsymbol{\gamma}$, the posterior distribution of o_d can be formulated as

$$p(o_d|\boldsymbol{\theta}_d,\boldsymbol{\pi}_d,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\gamma}) = \frac{p(\boldsymbol{\theta}_d,\boldsymbol{\pi}_d|o_d,\boldsymbol{\alpha},\boldsymbol{\gamma})\cdot p(o_d|\boldsymbol{\xi})}{\sum_{o_d=1}^{K}p(\boldsymbol{\theta}_d,\boldsymbol{\pi}_d|o_d,\boldsymbol{\alpha},\boldsymbol{\gamma})\cdot p(o_d|\boldsymbol{\xi})} \propto p(\boldsymbol{\theta}_d,\boldsymbol{\pi}_d|o_d,\boldsymbol{\alpha},\boldsymbol{\gamma})\cdot p(o_d|\boldsymbol{\xi}) \tag{14}$$

For efficiency, we only use the observed $\boldsymbol{\theta}_d$ and $\boldsymbol{\pi}_d$ to model and predict $\mathbf{o}$. Thus, the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ should be integrated out from Eq. (14). In addition, since the emotion-document distribution $\boldsymbol{\pi}_d$ is considered individually as a parameter during the modeling process, we do not consider the influence of the levels of emotional intensity for calculating topic-document distributions $\boldsymbol{\theta}_d$. That is, we apply

$$\theta_{t,d} = \frac{\sum_{s=1}^{S}(n^{d}_{t,s,d}+\alpha_{t,s,d})}{\sum_{t=1}^{T}\sum_{s=1}^{S}(n^{d}_{t,s,d}+\alpha_{t,s,d})}$$

to obtain the probability of the presence of the t-th topic in the d-th document. We now model openness to experience on the training corpus according with the formulation as

$$\begin{aligned} &p(o_d|\boldsymbol{\theta}_d,\boldsymbol{\pi}_d,\boldsymbol{\xi}) \\ &= \frac{\left(\iint_{\boldsymbol{\alpha},\boldsymbol{\gamma}} p(\boldsymbol{\theta}_d,\boldsymbol{\pi}_d|o_d,\boldsymbol{\alpha},\boldsymbol{\gamma})\cdot p(\boldsymbol{\alpha})\cdot p(\boldsymbol{\gamma})\cdot \mathrm{d}\boldsymbol{\alpha}\mathrm{d}\boldsymbol{\gamma}\right)\cdot p(o_d|\boldsymbol{\xi})}{\sum_{o_d=1}^{K} p(o_d|\boldsymbol{\xi})\cdot\left(\iint_{\boldsymbol{\alpha},\boldsymbol{\gamma}} p(\boldsymbol{\theta}_d,\boldsymbol{\pi}_d|o_d,\boldsymbol{\alpha},\boldsymbol{\gamma})\cdot p(\boldsymbol{\alpha})\cdot p(\boldsymbol{\gamma})\cdot \mathrm{d}\boldsymbol{\alpha}\mathrm{d}\boldsymbol{\gamma}\right)} \\ &\propto p(\boldsymbol{\theta}_d,\boldsymbol{\pi}_d|o_d)\cdot p(o_d|\boldsymbol{\xi}) = \prod_{t=1}^{T}p(\theta_{t,d}|o_d)\cdot\prod_{s=1}^{S}p(\pi_{s,d}|o_d)\cdot p(o_d|\boldsymbol{\xi}) \end{aligned} \tag{15}$$

When $p(\theta_{t,s,d}|o_d)$, $p(\pi_{s,d}|o_d)$, and $p(o_d|\boldsymbol{\xi})$ are obtained, we can then make use of them to predict the levels of openness to experience for each document in the testing corpus by the MAP estimation:

$$p(o_{d'}|\boldsymbol{\theta}_{d'},\boldsymbol{\pi}_{d'},\boldsymbol{\xi}) = \max_{o\in 1,\ldots,S}\prod_{t=1}^{T}p(\theta_{t,d'}|o)\cdot\prod_{s=1}^{S}p(\pi_{s,d'}|o)\cdot p(o|\boldsymbol{\xi}) \tag{16}$$

To effectively apply MAP on $p(\theta_{t,d}|o_d)$, $p(\pi_{s,d}|o_d)$, and $p(o_d|\boldsymbol{\xi})$ which are continuous in both training and testing corpora, we take the assumption that the values of $p(\theta_{t,d}|o_d)$ and $p(\pi_{s,d}|o_d)$ associated with o_d are distributed according to Normal distributions. Let $\hat{\mu}_{t,o}$ and $\tilde{\mu}_{s,o}$ be the means of the values in $\theta_{t,d}$ and $\pi_{s,d}$ associated with the openness to experience label o, respectively; let $\hat{\sigma}^2_{t,o}$ and $\tilde{\sigma}^2_{s,o}$ be the variances of the values in $\theta_{t,d}$ and $\pi_{s,d}$ associated with the openness to experience label o, respectively. Then, $p(\theta_{t,d'}|o)$ and $p(\pi_{s,d'}|o)$ can be computed by plugging $\theta_{t,d'}$ and $\pi_{s,d'}$ into the expressions for Normal distributions parameterized by $(\hat{\mu}_{t,o},\hat{\sigma}^2_{t,o})$ and $(\tilde{\mu}_{s,o},\tilde{\sigma}^2_{s,o})$, respectively. Thus far, we have

$$p(\theta_{t,d'}|o) \propto \frac{1}{\hat{\sigma}_{t,o}}\exp\left(-\frac{(\theta_{t,d'}-\hat{\mu}_{t,o})^2}{2\hat{\sigma}^2_{t,o}}\right), \tag{17}$$

$$p(\pi_{s,d'}|o) \propto \frac{1}{\tilde{\sigma}_{s,o}}\exp\left(-\frac{(\pi_{s,d'}-\tilde{\mu}_{s,o})^2}{2\tilde{\sigma}^2_{s,o}}\right), \tag{18}$$

and

$$p(o|\boldsymbol{\xi}) \propto D_o+\xi_o. \tag{19}$$

where Eq. (19) is derived in that the Dirichlet and the multinomial are a conjugate pair. Finally, the MAP prediction rule can be formulated in log-space as

$$p(o_{d'}|\boldsymbol{\theta}_{d'},\boldsymbol{\pi}_{d'},\boldsymbol{\xi})$$

$$\propto \min_{o\in 1,\dots,K} \sum_{t=1}^{T} \frac{\left(\theta_{t,d'} - \hat{\mu}_{t,o}\right)^2}{2\hat{\sigma}_{t,o}^2} + \sum_{t=1}^{T} \log \hat{\sigma}_{t,o} + \sum_{s=1}^{S} \frac{\left(\pi_{s,d'} - \tilde{\mu}_{s,o}\right)^2}{2\tilde{\sigma}_{s,o}^2} + \sum_{s=1}^{S} \log \tilde{\sigma}_{s,o} - \log\left(D_o + \xi_o\right) \quad (20)$$

Eq. (20) shows a data-driven prediction procedure, because it learns the correlations of $\mathbf{o}$ and $(\mathbf{\Theta}, \mathbf{\Pi})$ in terms of both structures of training and testing corpora. Compared with the procedures that impose some prior relationships between $\mathbf{o}$ and $(\mathbf{\Theta}, \mathbf{\Pi})$, our procedure may be more flexible on dealing with the dynamics of the UGCs in the context of big data.

6. Experiments and discussion

6.1. Datasets

In this section, experiments are conducted to validate the proposed method with two real-world UGC datasets. The first one is the stream-of-consciousness essay dataset (hereafter called essay dataset).[3] This corpus collected by Pennebaker and King (1999) originally contains 2468 passages with a total of 1.9 million words from different individuals whose openness to experience are assessed using a questionnaire and tagged with “low” or “high”. The second data is collected from myPersonality (hereafter called myPersonality dataset), which is a third-party Facebook application allowing users to take a set of psychological questionnaires[4]. The corpus used in our experiments is a subset of myPersonality users ($N = 250$), containing a total of 9917 status updates and the corresponding levels of openness to experience for the users (Kosinski, Stillwell, & Graepel, 2013). For essay dataset, the raw corpus is preprocessed by (a) modifying all letters to lowercase, (b) removing punctuation and stopwords, (c) stemming and lemmatizing to reduce morphological variation, and (d) removing useless words with respect to tf-idf value (10822 words remain in the vocabulary). In addition, we remove one sample from the corpus that only contains “Err:508”. Finally, the corpus used in our experiments contains 2467 passages with a total of 730,477 words. For myPersonality dataset, 9917 Facebook updates are first merged into 250 textual samples corresponding to the 250 users. The preprocessing procedure of myPersonality dataset is similar to that of essay dataset except feature and sample reduction: Because the size of the data accessible from myPersonality is small ($N = 250$), we do not remove any words with low tf-idf value from the corpus (13180 words remain in the vocabulary), but remove the users corresponding to less than 5 words. Finally, we collect 232 textual samples with a total of 77,318 words.

6.2. Experimental setting

Four representative classifiers, including the support vector machine (SVM), the naïve Bayes classifier (NB), the k-nearest neighbors classifier (kNN), and the recurrent neural networks (RNN), are selected as the compared methods in the experiments. Since the proposed method is built within the LDA-based framework, we also want to compare the performance of the proposed method with that of LDA. However, LDA is unsupervised and cannot directly predict openness to experience. We thus apply LDA as a feature extraction method and model the selected classifiers (i.e. SVM, NB, and kNN) using the topics extracted from the corpus by LDA. For RNN, we apply the long short-term memory (LSTM) for feature extraction and representation, and the softmax function to generate the final classification results. Therefore, there is a total of seven compared methods applied in our experiments, namely, SVM, NB, kNN, LDA+SVM, LDA+NB, LDA+kNN, and LSTM+softmax. We apply the linear kernel and the default setting $C = 0.99$ for SVM, and set the neighbor parameter $k = 3$ for kNN. For NB, we apply multinomial NB for word frequency modeling (i.e. modeling directly on the corpus), and Gaussian NB for topic preference modeling (i.e. modeling on the topic distributions from LDA). For LSTM, we set the hidden layer with the size of 128.

For LDA and the proposed method, we follow a line of works (e.g. Lin & He, 2009; Liu et al., 2016) to set all the hyperparameters $\boldsymbol{\alpha}$ equally as $\frac{50}{\#topic}$ and $\boldsymbol{\beta}$ equally as 0.01. For the hyperparameter $\boldsymbol{\xi}$ and $\boldsymbol{\gamma}$ in the proposed method, we set both equally to be 0.01 as we have no prior knowledge for their associated variables. We will pay specific attention to the hyperparameters of the emotion-document distributions and the topic-emotion-document distributions for the proposed method in Section 6.9 by empirically test the sensitivity of the proposed method with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$. It is also worth noting here that, Maximum Likelihood (ML) or MAP may be applied to estimate these hyperparameters in a data-driven manner (Lin & He, 2009). We leave the estimation of the hyperparameters of the proposed method in a more principled way as future work.

We use n-fold Cross-Validation (CV) to evaluate all the selected methods and report the average results in the experiments. For the proposed method and the LDA-based classification, $n - 1$ folds of the corpus are used to build the model, i.e. to estimate $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ for the proposed method and LDA, and $\boldsymbol{\pi}$ for the proposed method. The topics as well as the emotional intensity labels in the last fold will be then generated according to Eqs. (12) and (13) for the proposed method. For LDA, only the topic will be sampled and generated according to

$$\phi^{\text{train}}_{w_i,z_i} \cdot (n^{d^{\text{test}},\neg i}_{z_i} + \alpha_{z_i}) \quad (21)$$

Finally, the generated topics (and the emotional intensity labels) in the testing corpus are used to conduct the classification task. 5-fold CV is applied for classification tasks on essay dataset and 3-fold CV is applied on myPersonality dataset because of its small scale.

6.3. Convergence analysis of the proposed method

To determine the number of iterations required for the Gibbs sampler to generate stable inference results, we run the proposed method and LDA on essay dataset for convergence analysis where Gibbs sampling will be iterated 1500 times. The minus log likelihood of the models versus the iteration time are reported in Fig. 5,[5] where the proposed method is named as a Valence (emotion)-Topic-Openness mixture model for openness Prediction (VTOP, hereafter *the proposed method* and *VTOP* are used interchangeably).

It can be seen from Fig. 5(a) that, the burn-in period of LDA is limited within 100 iterations of Gibbs sampling, while VTOP does not fully converge even after 1000 iterations (see Fig. 5(b)). For tradeoff, samples generated after 100 iterations are used as the results of LDA, and those after 1000 iterations are selected as the results of VTOP, in our experiments.

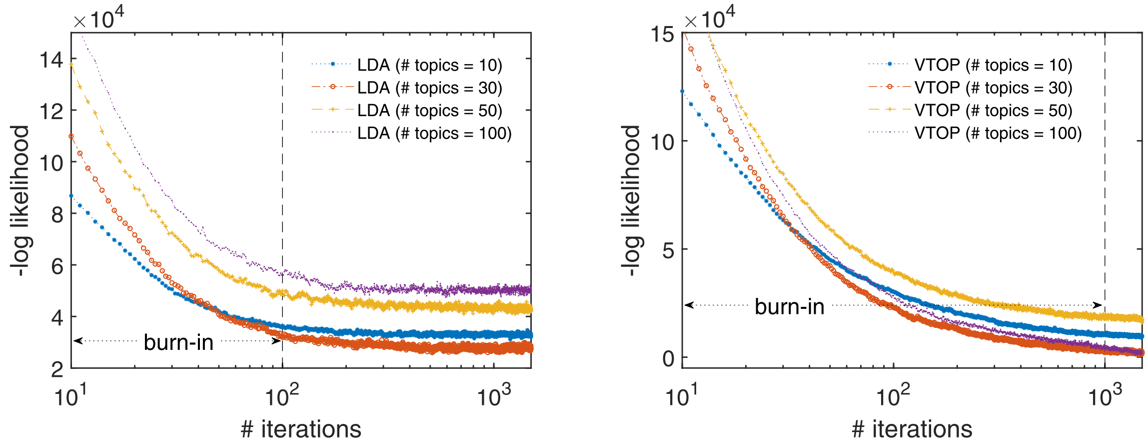
6.4. Evaluation metrics

Five performance evaluation metrics, namely accuracy, precision (also called positive predictive value), recall (also called

[3] https://github.com/SenticNet/personality-detection.

[4] http://www.mypersonality.org/wiki.

[5] For clear representation, we slightly shift some of the curves vertically in Fig. 5 to avoid overlapping.



(a) LDA (b) VTOP

Fig. 5. Convergence analysis of Gibbs sampling for LDA and VTOP on essay dataset.

sensitivity), F_1 score, and Receiver Operating Characteristic (ROC) curve, are applied to get full evaluations of the results obtained by VTOP and the compared methods. For 0–1 classification tasks, the classification results consist of four cases, denoting as *true positives* (TP, predicted 1 and actually 1), *true negatives* (TN, predicted 0 and actually 0), *false positives* (FP, predicted 1 but actually 0), and *false negatives* (FN, predicted 0 but actually 1). The former four metrics are defined as follows:

$$accuracy = \frac{\#TP + \#TN}{\#TP + \#TN + \#FP + \#FN},$$

$$precision = \frac{\#TP}{\#TP + \#FP},$$

$$recall = \frac{\#TP}{\#TP + \#FN},$$

$$F_1\ score = \frac{2 \cdot precision \cdot recall}{precision + recall}.$$

With the above definitions, accuracy is the proportion of true results (both TP and TN) among the total number of cases examined. Precision is the proportion of TP among the total number of cases predicted as positive. Recall is the proportion of TP among the total number of cases actually positive. F_1 score is the harmonic average of precision and recall. The last metric, ROC curves, is created by plotting the TP rate versus the FP rate at different threshold settings, to illustrate the diagnostic ability of a binary classifier. We apply the normalized Area Under the Curve (AUC) as the ROC index in our experiments.

6.5. Results and discussion

The prediction results for all selected methods on the two datasets are shown in Tables 3 and 4, respectively, where the results reported for LDA+SVM, LDA+NB, LDA+*k*NN, LSTM+softmax, and VTOP are their best results that may correspond to different numbers of topics. Value in bold in Tables 3 and 4 means that it is highest among the values in the same column. To further investigate the influence of the number of topics on the prediction ability of LDA and VTOP, we report the prediction results with respect to the number of topics in Figs. 6 and 7, respectively, where the shaded area around the line denotes the standard error.

As exhibited in Table 3, the proposed method achieves the best performance among the selected methods in terms of accuracy, precision, recall, F_1 score, and ROC (AUC). LSTM+softmax outperforms most of the selected methods except the proposed method. In addition, the results also show that LDA-based methods including VTOP perform significant better than SVM and *kNN* in most of cases, indicating that topic-based modeling is more effective than word-frequency-based modeling in predicting openness to experience. We may also find that LDA+SVM achieves slightly inferior results than SVM in terms of recall, F_1 score, and ROC (AUC), although it holds higher results in terms of accuracy and precision. This implies that the topics extracted by LDA may lead to classification bias for SVM which is very sensitive to parameters and data structure.

For myPersonality dataset, the proposed method again outperforms all other compared methods in terms of all the selected evaluation metrics, as shown in Table 4. We may find that the ROC (AUC) values for the selected methods are all approximately 50 except VTOP, illustrating that they are nearly uninformative classifiers for openness to experience prediction on such a small-scale corpus. By contrast, VTOP performs superiorly with an AUC of 60.85, indicating that VTOP can effectively predict

Table 3
Best performance of the selected methods on essay dataset (%).

	accuracy	precision	recall	F_1 score	ROC (AUC)
SVM	54.94	54.94	55.96	54.89	55.28
LDA + SVM	58.66	57.80	55.20	49.63	55.20
NB	57.29	57.26	57.11	56.96	63.10
LDA + NB	60.22	60.19	59.94	59.73	59.65
*k*NN	53.40	53.55	53.49	53.27	54.14
LDA + *k*NN	57.30	57.04	56.88	56.71	59.10
LSTM + softmax	59.02	59.20	59.16	59.18	61.45
VTOP	**62.67**	**62.58**	**62.24**	**61.95**	**64.02**

Table 4
Best performance of the selected methods on myPersonality dataset (%).

	Accuracy	Precision	Recall	F_1 score	ROC (AUC)
SVM	58.57	59.65	58.57	57.23	49.25
LDA + SVM	69.40	49.30	69.94	57.73	50.00
NB	64.51	57.81	64.51	57.86	48.62
LDA + NB	64.34	55.78	64.35	58.29	51.61
*k*NN	56.88	62.61	56.88	56.12	47.53
LDA + *k*NN	63.99	60.74	63.99	60.52	50.86
LSTM + softmax	69.57	62.99	**70.44**	66.51	58.39
VTOP	**70.18**	**66.81**	70.18	**66.81**	**60.85**

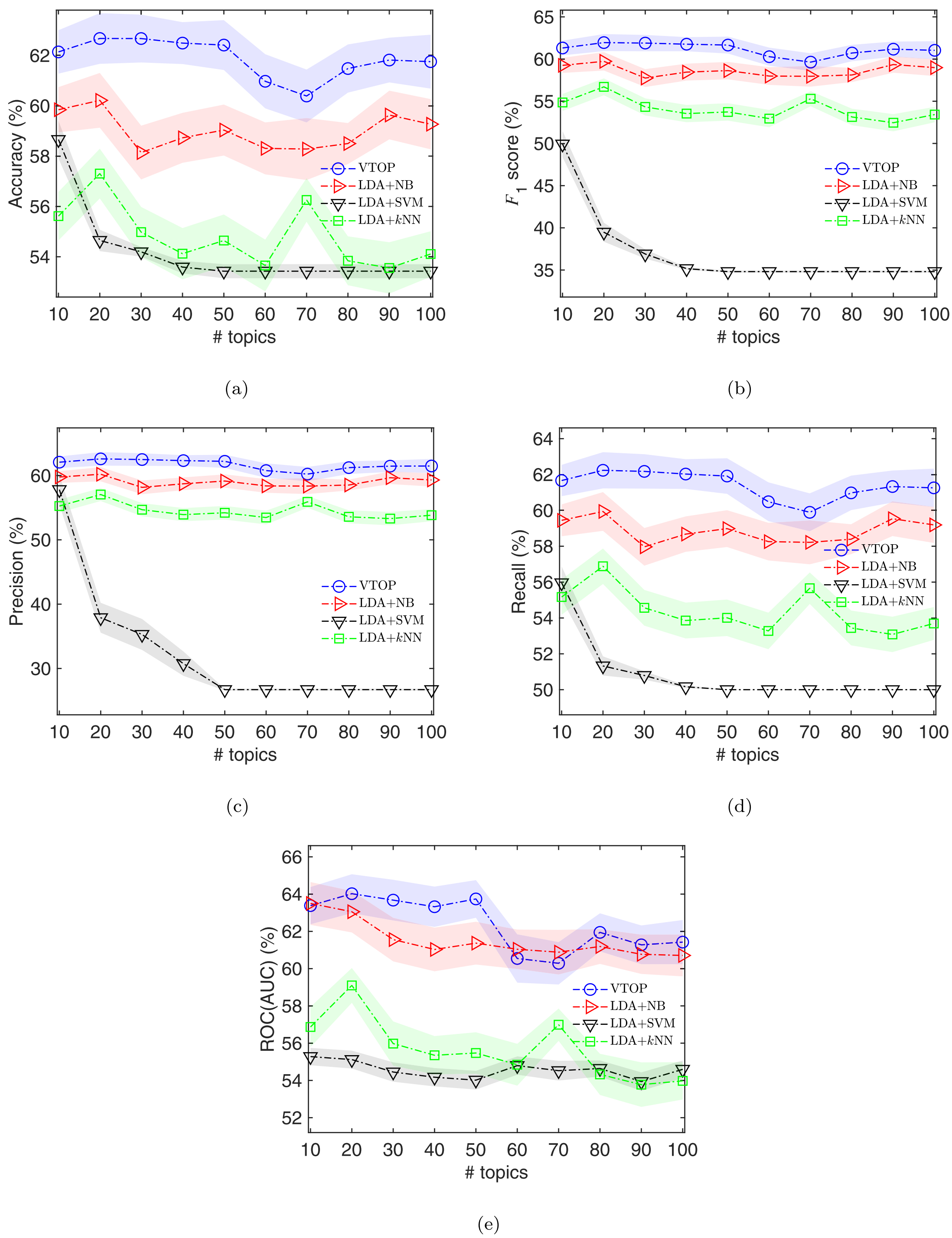


Fig. 6. Prediction results w.r.t. # topics for the selected methods on essay dataset, including the results of (a) accuracy, (b) precision, (c) recall, (d) F_1 score, and (e) ROC (AUC). The shaded area around the line denotes the standard error.

openness to experience with the information of topic and the emotional intensity even on a small-scale corpus. It is also noting that the performance of LSTM+softmax is similar (even better in recall) to VTOP, showing attractive potential of deep learning in dealing with textual data despite its less interpretable characteristic.

The results shown in Figs. 6 and 7 again verify the superiority of VTOP in most of cases, particularly on essay dataset. Specifically, VTOP achieves its best performance with 20 topics in terms of all the evaluation metrics on both two datasets. However, we may find that LDA+SVM achieves superior and stable performance regardless of the number of topics (see Fig. 7(a) and (d), while owns very poor results in terms of precision and ROC (AUC). This lies in the fact that myPersonaly dataset is an imbalanced corpus and LDA+SVM achieves extreme biased prediction results on such a dataset. By contrast, VTOP achieves balanced results in terms of all the evaluation metrics, indicating its superior ability to deal with the imbalanced dataset.

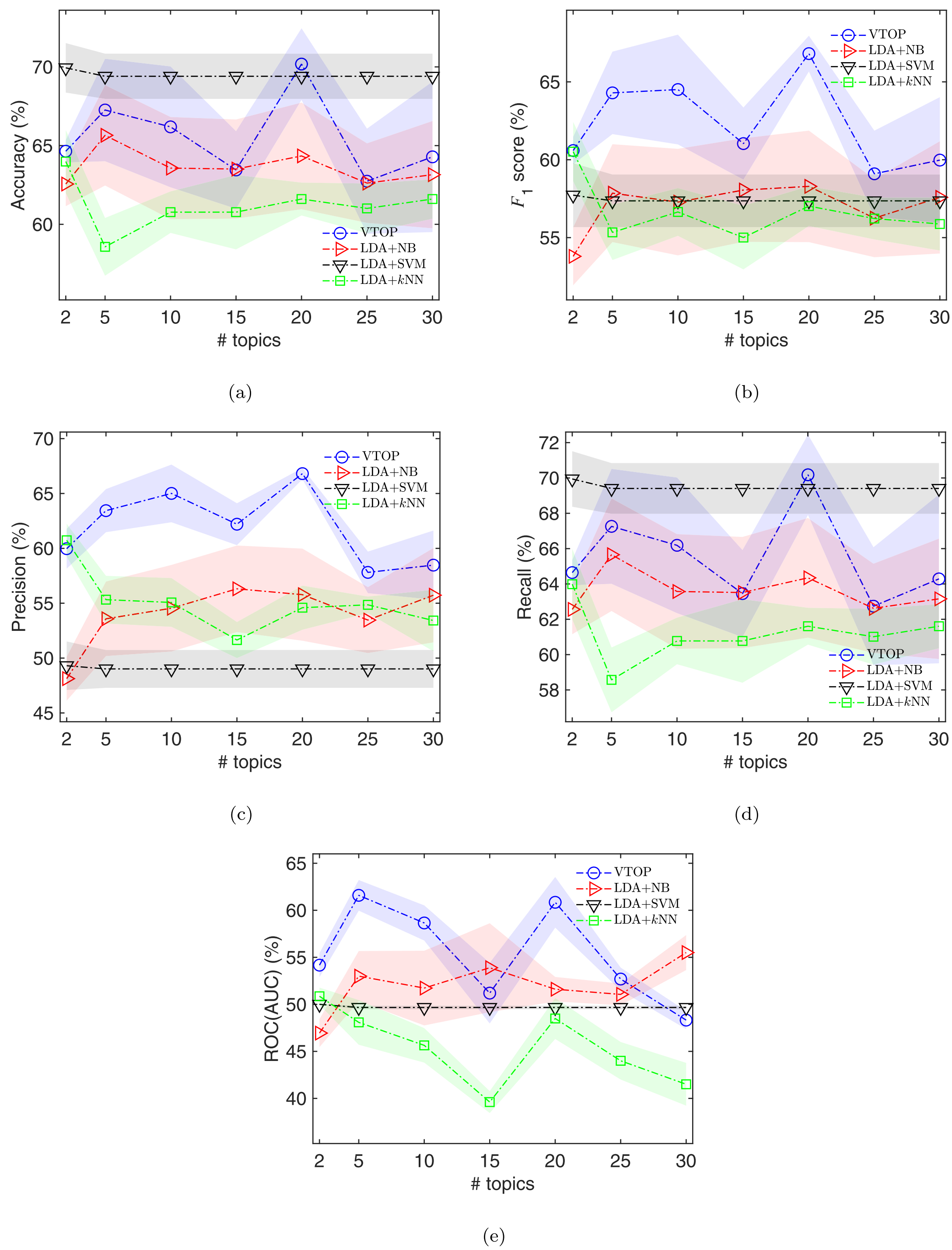


Fig. 7. Prediction results w.r.t. # topics for the selected methods on myPersonality dataset, including the results of (a) accuracy, (b) precision, (c) recall, (d) F_1 score, and (e) ROC (AUC). The shaded area around the line denotes the standard error.

6.6. Model interpretability analysis

Conceptually, the key difference between our method and the compared machine learning methods is the consideration of two interpretable variables (i.e., users' topic preference and expressed emotional intensity, as detailed earlier). These two variables are known as critical factors for constructing interpretable models that can help demonstrate why a particular level of openness to experience is predicted for a user. Methodologically, our method features two key novelties: a) Utilizing a generative learning framework (i.e., LDA) to extract topic preference and expressed emotional intensity by considering their causal relationship revealed by the existing empirical studies, and b) learning users' openness to experience by appropriately integrating topic preference and expressed emotional intensity.

Fig. 8 gives an instance that shows openness to experience level vs. users' topic preferences and expressed emotional intensity on myPersonality dataset (where five topics and two

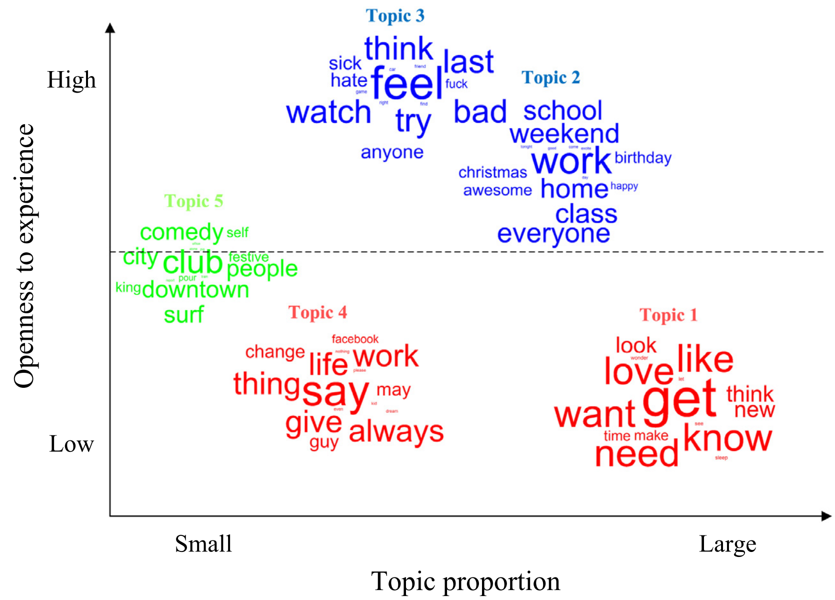


Fig. 8. Interpretability analysis on myPersonality dataset.

expressed emotional intensity levels are generated, and "topic proportion" in Fig. 8 means the proportion of the topic in the corpus). It can be observed that, topic 1 and topic 4 corresponding to low level of openness to experience contain a large proportion of self-related words (like "need" and "know"), while topic 2 and topic 3 corresponding to high level of openness to experience contain a large proportion of objective words (like "weekend" and "'everyone'). In addition, topic 2 and topic 3 corresponding to high level of openness to experience tend to include more emotional words like "hate", "happy", and "awesome" (emotional words also differentiate topic 3 and topics 4 and 5 which all contains a certain proportion of self-related words).

Altogether, Fig. 8 shows the discriminative ability of topic preference and expressed emotional intensity. It reveals that our method can effectively extract these two interpretable variables from the corpus and can effectively utilize them to automatically predict openness to experience. Compared with traditional "black box" machine learning methods, the proposed method seems more proper for some empirical studies (e.g., data-driven exploration in social science) where both comprehensibility and data processing ability play important roles. In addition, in spite of the effective prediction performance, the proposed method is particularly helpful for decision support in real-world applications due to its interpretability.

6.7. Efficiency analysis for VTOP on essay dataset

Section 6.3 analyzes the convergence of the proposed method and LDA, indicating a longer burn-in time of the proposed method. In order to analyze in detail the efficiency of the proposed method, we show the prediction accuracy with respect to the Gibbs sampling iterations for both the proposed method and LDA in Fig. 9, on essay dataset. It can be observed that, within the first 50 iterations, the proposed method fails in both the prediction accuracy and the increasing rate of the accuracy (i.e., the accuracy–iteration ratio), showing a relatively lower efficiency comparing with LDA. This is consistent with the results from the convergence analysis in Section 6.3. As the number of iterations grows up (after 50 iterations), the performance of the proposed method increases with the highest increasing rate, while that of LDA reaches its upper limit.

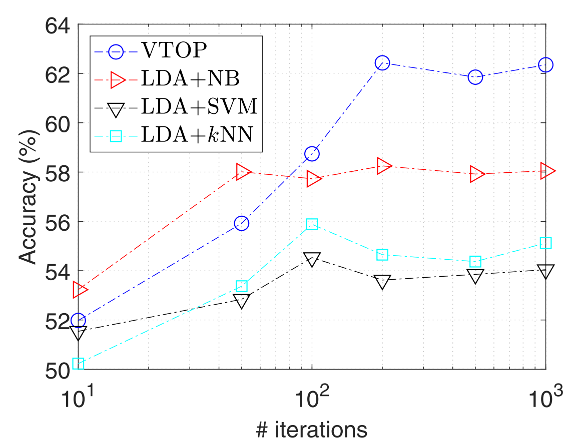


Fig. 9. Accuracy vs. iteration for LDA + NB, LDA + SVM, LDA + *k*NN, and the proposed method on essay dataset.

6.8. Discriminative ability analysis for topic distribution

In this section, we investigate the discriminative ability with respect to different numbers of topics extracted by the proposed method on essay dataset. Specifically, we measure and compare the differences of the topic distributions over the passages corresponding to $o = 0$ and $o = 1$ on training (four folds in CV for model training) and testing (the rest fold in CV for model testing) datasets, respectively. We apply the Kullback–Leibler divergence (KL-divergence) to measure these differences, as shown in Table 5 and Fig. 10. For clarity, we only show in Fig. 10 the topic-openness distributions with # topics $= 10, 30, 50,$ and 100, respectively. The KL-divergence between two distributions $\mathbf{p}$ and $\mathbf{q}$ is defined as

$$\mathrm{KL}(\mathbf{p}||\mathbf{q}) = \sum_i p_i \log \frac{p_i}{q_i},$$

indicating the extent to which $\mathbf{p}$ differ from $\mathbf{q}$. As shown in Table 5, the KL-divergence for word distributions is 0.278 for the training dataset and 0.276 for the testing dataset, both significantly higher than those for topic distributions. This implies a fact that all potential knowledge for openness to experience prediction is hidden among the words. However, as illustrated before, word fre-

Table 5
The Kullback–Leibler (KL) divergence between the distributions over words and topics for openness on essay dataset.

	All words	# topics									
		10	20	30	40	50	60	70	80	90	100
Training data	0.278	0.149	0.053	0.077	0.071	0.046	0.035	0.028	0.041	0.039	0.038
Testing data	0.276	0.071	0.038	0.031	0.024	0.016	0.016	0.015	0.020	0.019	0.010

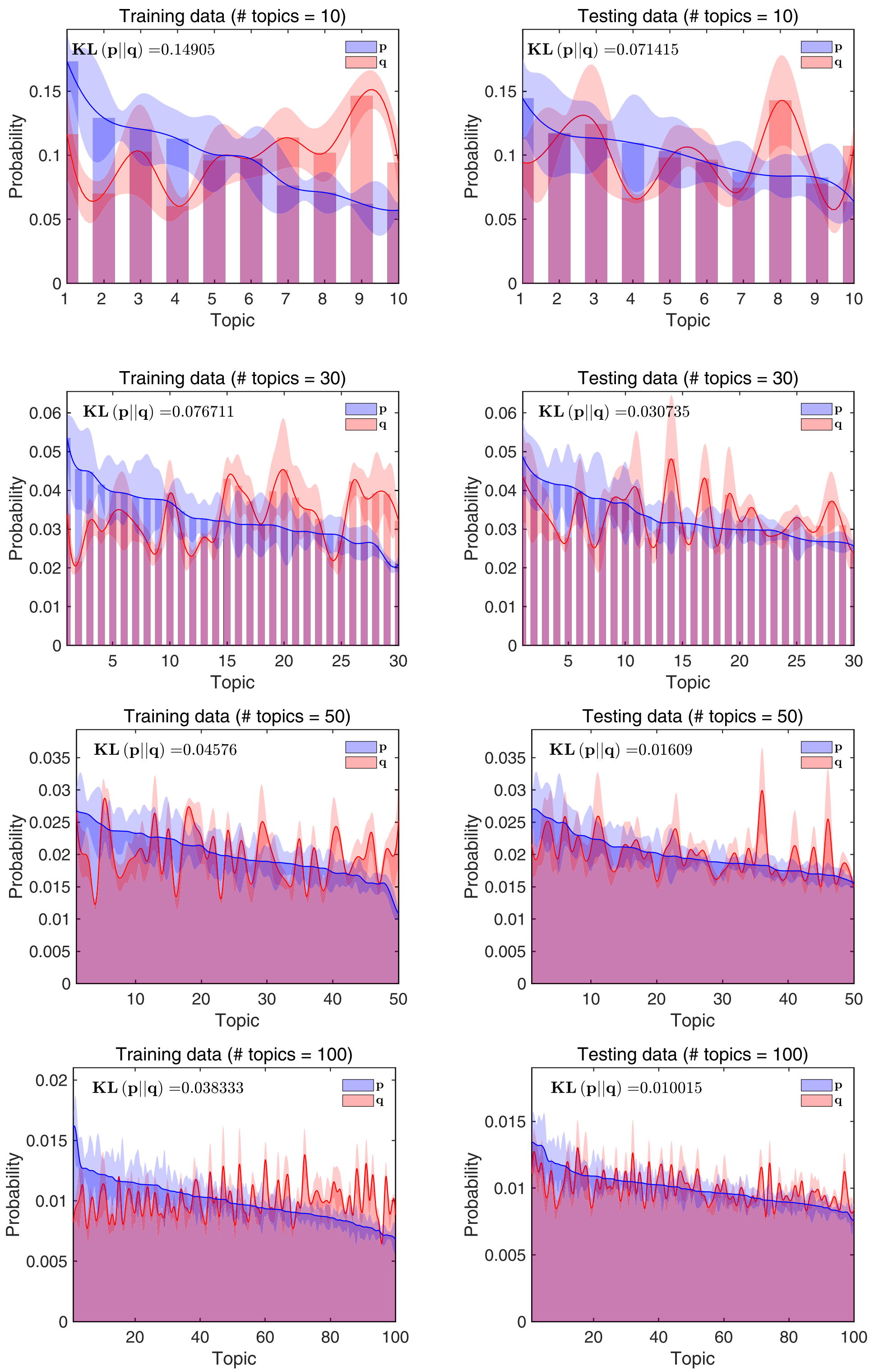


Fig. 10. The distributions over topics on the training (left) and testing (right) essay datasets with 10, 30, 50, and 100 topics, respectively, where topic indices are assigned according to the descending order of $p(\mathbf{z}|o=0)$. For convenience, denote $\mathbf{p} = p(\mathbf{z}|o=0)$ and $\mathbf{q} = p(\mathbf{z}|o=1)$. The shaded area around the line denotes the standard error of the distribution.

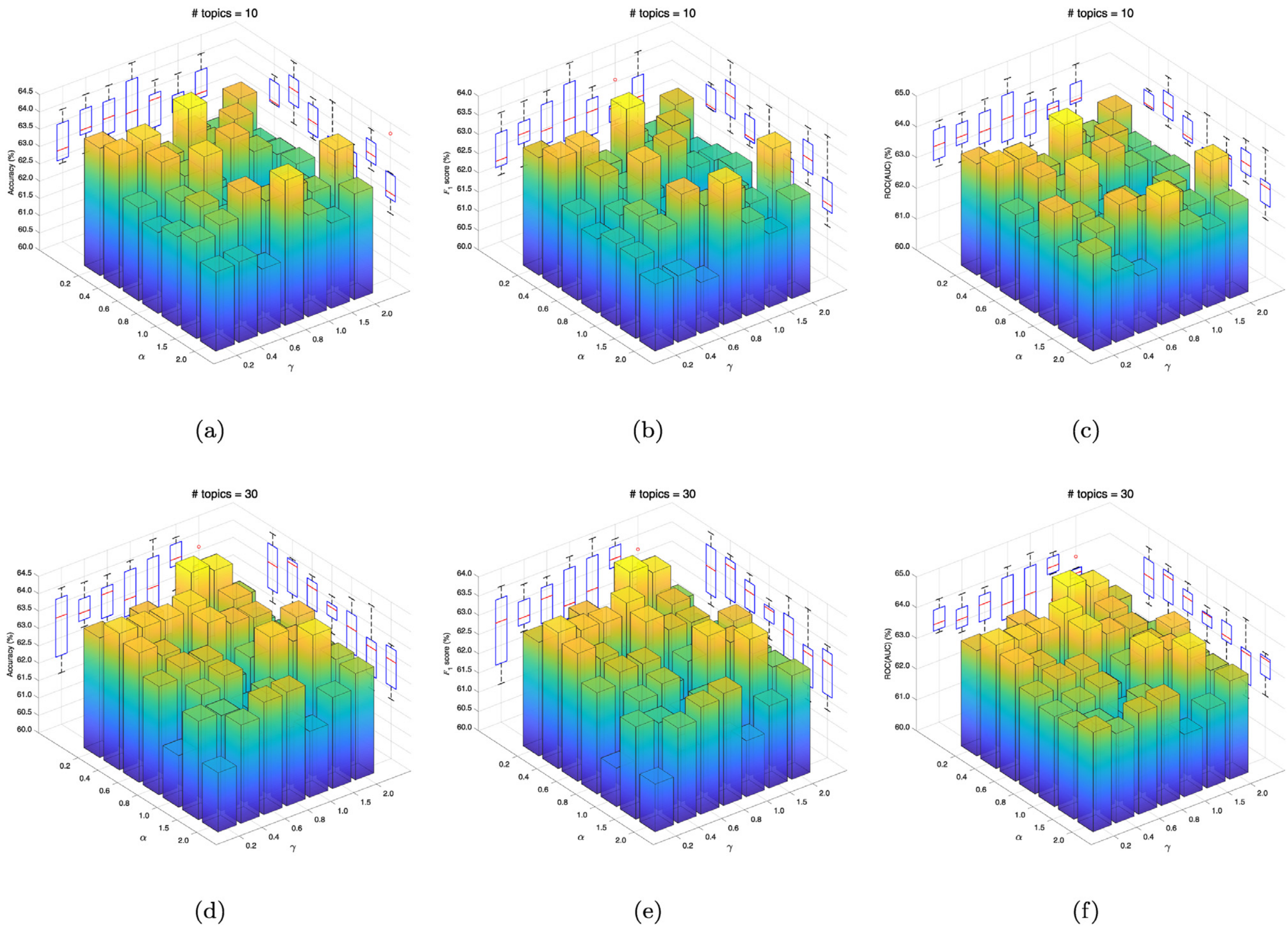


Fig. 11. Prediction results with $\alpha = \{0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0\}$ and $\gamma = \{0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0\}$, where the box-and-whisker plots shown in the planes of $\alpha = 0$ and $\gamma = 2.5$ depict the groups of results through their quartiles with a certain value of γ and α, respectively.

quency is a poor predictive variable for openness to experience on essay dataset, indicating that sparseness and high-dimensionality of words severely impairs potential knowledge discovery for openness to experience prediction. When the dimensionality has been dramatically reduced from 10,822 words to less than 100 topics, the differences of the distributions can be easily observed as revealed in Fig. 10, illustrating the strong discriminative power of the topics extracted by VTOP. However, such differences become indistinct when the number of extracted topics increases (from the top to the bottom in Fig. 10, consistent with the prediction performance shown in Fig. 6(a) and (e). This is possibly because the excessive topics exhibit redundancy.

6.9. Parameter tuning for VTOP on essay dataset

As observed from Figs. 6 and 7, we know that the number of topics play an important role in the proposed method. Superior performance is more likely to appear when the number of topics is less than 50 on essay dataset and 20 on myPersonality dataset. In this section, we focus on the hyperparameters of the model, i.e. the hyperparameter of the topic-emotion-document distributions (α), and the hyperparameter of the emotion-document distributions (γ). Specifically, we conduct parameter tuning under two scenarios in which # topics is set to be 10 and 30, respectively. Under each scenario, we do grid search with $\alpha = \{0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0\}$ and $\gamma = \{0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0\}$. The whole process requires $2 \times 7 \times 7$ experiments to investigate in detail the sensitivity of VTOP to different values of the parameters. The corresponding results are shown in Fig. 11, where only the results of accuracy, F_1 score, and ROC (AUC) are exhibited to keep clarity.

As seen from the parameter tuning results in Fig. 11, we may conclude that no distinct rules exist for the influence of these hyperparameters on model performance. However, there seems to be a tendency for VTOP to perform better with lower α and higher γ when the number of topics is 30, indicating a moderating role of the number of topics in the influence of α and γ on prediction results obtained by the proposed method. On the whole, most of the results in terms of all the evaluation metrics are in the range of [61,65], indicating a relatively stable and balanced performance for the proposed method.

7. Conclusions

In this study, we proposed a Bayesian topic-emotion-openness mixture model for modeling and predicting individuals' openness to experience from the textual data. It extends the latent Dirichlet allocation (LDA) to overcome the problem of high dimensionality and sparseness in modeling the textual data, and can effectively extract the topic preference and the expressed emotional intensity to fully understand individuals' openness to experience. We applied Gibbs sampling to obtain the estimated topic-emotion-document distributions $\boldsymbol{\theta}$ and emotion-document distributions $\boldsymbol{\pi}$, and applied the MAP estimation for modeling and predicting openness to experience using the extracted topics and emotional intensity labels in a data-driven manner. To verifying the effectiveness of the proposed method, we selected SVM, NB, *k*NN, their combinations with LDA, and the deep neural networks RNN, as the compared methods, and conducted extensive experiments on essay dataset and myPersonality dataset. In order to get comprehensive evaluations, we applied accuracy, precision, recall, and ROC (AUC) to evaluate the classification results of the selected methods. Evaluation results showed that the proposed method is more effective and stable than the compared methods, and also validated the predictive ability of topic preference and expressed emotional intensity which are indicated to have an effect on individuals' openness to experience in psychological literature. Compared with traditional "black box" intelligent methods, the proposed method seems more proper for the data-driven exploration in social science where both comprehensibility and data processing ability play important roles. In addition, in spite of the effective prediction performance, the proposed method is particularly helpful for decision support in real-world applications due to its interpretability.

Our study has several limitations. Firstly, although the proposed method provides interpretability for the empirical researchers and practitioners, performing inference with very large-scale datasets is a challenge due to its high complexity. Secondly, there are no sufficient public data sources for modeling and predicting individual characteristics like openness to experience. To our knowledge, the corpora used in this study, i.e. essay dataset and myPersonality dataset, are only the currently available datasets that can be applied in individual characteristics prediction. Thirdly, the proposed method only considers the topics and the emotional cues in the textual data for modeling and predicting openness to experience. Other potential predictive variables are not considered in the proposed method. Last but not least, the proposed method uses the MAP estimation by assuming that θ and π are distributed according to Normal distributions, whereas θ and π are also assumed to be Dirichlet distributed when extracting the topics and the emotional intensity labels. Such a discrepancy impairs the performance of the proposed method. We will focus on the above-mentioned issues in future. In addition, the proposed method focuses only on predicting openness to experience, while the extracted topic preference and expressed emotional intensity may be potentially influential to other personality traits. How to help improve the learning of other personality traits using the knowledge of topic preference and emotional intensity extracted for openness to experience, i.e., an effective transfer learning procedure, will be focused in our future work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Credit authorship contribution statement

Yishi Zhang: Conceptualization, Writing - original draft, Supervision, Formal analysis. **Haiying Wei:** Formal analysis. **Yaxuan Ran:** Data curation. **Yang Deng:** Writing - review & editing. **Dan Liu:** Writing - review & editing.

Acknowledgments

We appreciate the editor and two anonymous reviewers for their constructive comments and suggestions. This study was supported in part by National Natural Science Foundation of China under grants 71702066, 71701218, 71802192, 71772077, and in part by National Social Science Foundation under grant 16CTQ022.

References

Alam, F., Stepanov, E. A., & Riccardi, G. (2013). Personality traits recognition on social network - Facebook. In *Seventh international aaai conference on weblogs and social media (ICWSM)* (pp. 5–8).

Araque, O., Corcuera-Platas, I., Snchez-Rada, J. F., & Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications, 77*, 236–246. doi:10.1016/j.eswa.2017.02.002. http://www.sciencedirect.com/science/article/pii/S0957417417300751.

Baer, M., & Oldham, G. R. (2006). The curvilinear relation between experienced creative time pressure and creativity: Moderating effects of openness to experience and support for creativity. *Journal of Applied Psychology, 91*(4), 963–970. doi:10.1037/0021-9010.91.4.963.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Leanrning Research, 3*(4–5), 993–1022.

Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science, 35*(6), 953–975.

Christensen, A. P., Cotter, K. N., & Silvia, P. J. (2018). Reopening openness to experience: A network analysis of four openness to experience inventories. *Journal of Personality Assessment*, In–Press. doi:10.1080/00223891.2018.1467428.

Correa, T., Hinsley, A. W., & de Zúñiga, H. G. (2010). Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in Human Behavior, 26*(2), 247–253. doi:10.1016/j.chb.2009.09.003.

DeYoung, C. G., Grazioplene, R. G., & Peterson, J. B. (2012). From madness to genius: The openness/intellect trait domain as a paradoxical simplex. *Journal of Research in Personality, 46*(1), 63–78. doi:10.1016/j.jrp.2011.12.003.

Do, H. H., Prasad, P., Maag, A., & Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: A comparative review. *Expert Systems with Applications, 118*, 272–299. doi:10.1016/j.eswa.2018.10.003. http://www.sciencedirect.com/science/article/pii/S0957417418306456

Farnadi, G., Zoghbi, S., Moens, M., & Cock, M. D. (2013). Recognising personality traits using facebook status updates. *Workshop on computational personality recognition (wcpr13) in international aaai conference on weblogs and social media (icwsm13)*. doi:10.1177/0269216314552091.

Gers, F. A., Schraudolph, N. N., & Schmidhuber, J. (2003). Learning precise timing with lstm recurrent networks. *Journal of Machine Learning Research, 3*(1), 115–143.

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011a). Predicting personality from twitter. In *2011 ieee third international conference on privacy, security, risk and trust and 2011 ieee third international conference on social computing* (pp. 149–156). IEEE. doi:10.1109/PASSAT/SocialCom.2011.33.

Golbeck, J., Robles, C., & Turner, K. (2011b). Predicting personality with social media. In *2011 annual conference extended abstracts on human factors in computing systems* (pp. 253–262). New York, New York, USA: ACM Press. doi:10.1145/1979742.1979614.

Hasan, S., Ferguson, J.-P., & Koning, R. (2015). The lives and deaths of jobs: Technical interdependence and survival in a job structure. *Organization Science, 26*(6), 1665–1681. doi:10.1287/orsc.2015.1014.

Holtgraves, T. (2011). Text messaging, personality, and the social context. *Journal of Research in Personality, 45*(1), 92–99. doi:10.1016/j.jrp.2010.11.015.

Humphreys, A., & Wang, R. J.-H. (2018). Automated text analysis for consumer research. *Journal of Consumer Research, 44*(6), 1274–1306. doi:10.1093/jcr/ucx104.

Iacobelli, F., Gill, A. J., Nowson, S., & Oberlander, J. (2011). Large scale personality classification of bloggers. In *Proceedings of the 4th international conference on affective computing and intelligent interaction*. In *ACII'11* (pp. 568–577). Berlin, Heidelberg: Springer-Verlag.

Islam, J. U., Rahman, Z., & Hollebeek, L. D. (2017). Personality factors as predictors of online consumer engagement: An empirical investigation. *Marketing Intelligence and Planning, 35*(4), 510–528. doi:10.1108/MIP-10-2016-0193.

Jo, Y., & Oh, A. (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of the 4th acm international conference on web search and data mining* (pp. 815–824). New York, NY, USA: ACM.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, 110*(15), 5802–5805. doi:10.1073/pnas.1218772110.

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th acm conference on information and knowledge management* (pp. 375–384). New York, NY, USA: ACM.

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue, 16*(3), 31–57. doi:10.1145/3236386.3241340.

Liu, Y., Wang, J., & Jiang, Y. (2016). PT-LDA: A latent variable model to predict personality traits of social network users. *Neurocomputing, 210*, 155–163. doi:10.1016/j.neucom.2015.10.144.

Liu, Y., Wang, J., Jiang, Y., Sun, J., & Shang, J. (2018). Identifying impact of intrinsic factors on topic preferences in online social media: A nonparametric hierarchical bayesian approach. *Information Sciences, 423*, 219–234. doi:10.1016/j.ins.2017.09.041.

Ma, Y., Chen, G., & Wei, Q. (2017). Finding users preferences from large-scale online reviews for personalized recommendation. *Electronic Commerce Research, 17*(1), 3–29. doi:10.1007/s10660-016-9240-9.

Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in Neural Information Processing Systems* (pp. 121–128). Curran Associates, Inc..

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*(2), 175–215. doi:10.1111/j.1467-6494.1992.tb00970.x.

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology, 90*(5), 862–877. doi:10.1037/0022-3514.90.5.862.

Mulyanegara, R. C., Tsarenko, Y., & Anderson, A. (2009). The Big Five and brand personality: Investigating the impact of consumer personality on preferences towards particular brand personality. *Journal of Brand Management, 16*(4), 234–247. doi:10.1057/palgrave.bm.2550093.

Nallapati, R. M., Ahmed, A., Xing, E. P., & Cohen, W. W. (2008). Joint latent topic models for text and citations. In *Proceeding of the 14th acm sigkdd international conference on knowledge discovery and data mining* (pp. 542–550). New York, New York, USA: ACM Press. doi:10.1145/1401890.1401957.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., … Seligman, M. E. P. (2015). Automatic personality assessment through social media language.. *Journal of Personality and Social Psychology, 108*(6), 934–952. doi:10.1037/pspp0000020.

Pennebaker, J. W., Chung, C., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007. *The University of Texas at Austin*.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296–1312. doi:10.1037/0022-3514.77.6.1296.

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 248–256). doi:10.3115/1699510.1699543.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., & Steyvers, M. (2010). Learning author-topic models from text corpora. In *Acm transactions on information systems: 28* (pp. 1–38). doi:10.1145/1658377.1658381.

Rubin, T. N., Chambers, A., Smyth, P., & Steyvers, M. (2012). Statistical topic models for multi-label document classification. *Machine Learning, 88*(1–2), 157–208. doi:10.1007/s10994-011-5272-5.

Rudd, M., Hildebrand, C., & Vhos, K. D. (2018). Inspired to create: Awe enhances openness to learning and the desire for experiential creation. *Journal of Marketing Research*, In–Press.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one, 8*(9), e73791. doi:10.1371/journal.pone.0073791.

Shi, Z., Lee, G. M., & Whinston, A. B. (2016). Toward a better understanding of business proximity: Topic modeling for industry intelligence. *MIS Quarterly, 40*(4), 1035–1056.

Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin, 134*(1), 138–161. doi:10.1037/0033-2909.134.1.138.

Subramanian, R., Wache, J., Abadi, M. K., Vieriu, R. L., Winkler, S., & Sebe, N. (2018). ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing, 9*(2), 147–160. doi:10.1109/TAFFC.2016.2625250.

Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012). Predicting dark triad personality traits from Twitter usage and a linguistic analysis of tweets. In *11th international conference on machine learning and applications: 2* (pp. 386–393). IEEE. doi:10.1109/ICMLA.2012.218.

Tandera, T., Hendro, Suhartono, D., Wongso, R., & Prasetio, Y. L. (2017). Personality prediction system from Facebook users. *Procedia Computer Science*. doi:10.1016/j.procs.2017.10.016.

Tok, S., Koyuncu, M., Dural, S., & Catikkas, F. (2010). Evaluation of International Affective Picture System (IAPS) ratings in an athlete population and its relations to personality. *Personality and Individual Differences, 49*(5), 461–466. doi:10.1016/j.paid.2010.04.020.

Wald, R., Khoshgoftaar, T., & Sumner, C. (2012). Machine prediction of personality from Facebook profiles. In *Ieee 13th international conference on information reuse & integration (iri)* (pp. 109–115). IEEE. doi:10.1109/IRI.2012.6302998.

Wang, H., Wu, F., Lu, W., Yang, Y., Li, X., Li, X., & Zhuang, Y. (2018). Identifying objective and subjective words via topic modeling. *IEEE Transactions on Neural Networks and Learning Systems, 29*(3), 718–730. doi:10.1109/TNNLS.2016.2626379.

Wang, X., & McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining* (pp. 424–433). New York, New York, USA: ACM Press. doi:10.1145/1150402.1150450.

Wang, Y., & Chaudhry, A. (2018). When and how managers' responses to online reviews affect subsequent reviews. *Journal of Marketing Research, LV*, 163–177.

Weller, J., Ceschi, A., Hirsch, L., Sartori, R., & Geisler, M. (2018). Accounting for individual differences in decision-making competence: Personality and gender differences. *Frontiers in Psychology, 9*, 1–13. doi:10.3389/fpsyg.2018.02258.

Woo, S. E., Saef, R., & Parrigon, S. (2015). Openness to experience. In *International encyclopedia of the social & behavioral sciences* (pp. 231–235). Elsevier. doi:10.1016/B978-0-08-097086-8.25072-1.

Zabkar, V., Arslanagic-Kalajdzic, M., Diamantopoulos, A., & Florack, A. (2017). Brothers in blood, yet strangers to global brand purchase: A four-country study of the role of consumer personality. *Journal of Business Research, 80*(May), 228–235. doi:10.1016/j.jbusres.2017.06.006.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, 818–833.

Zhu, J., Ahmed, A., & Xing, E. (2012). MedLDA: Maximum margin supervised topic models. *Journal of Machine Learning Research, 13*, 2237–2278. doi:10.1145/1553374.1553535.