Expert Systems With Applications 114 (2018) 210–223

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews

Lu-yu Dong[a], Shu-juan Ji[b,c,*], Chun-jin Zhang[d], Qi Zhang[a], DicksonK.W. Chiu[e], Li-qing Qiu[a], Da Li[a]

[a] College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao, P.R. China
[b] Key Laboratory for Wisdom Mine Information Technology of Shandong Province, Shandong University of Science and Technology, Qingdao, P.R. China
[c] Shandong Provincial Key Laboratory of Novel Distributed Computer Software Technology, Shandong Normal University, Jinan, P. R. China
[d] Network Information Center(NIC), Shandong University of Science and Technology, Qingdao, P.R. China
[e] Faculty of Education, The University of Hong Kong, Hong Kong, P.R. China

ARTICLE INFO

Article history:
Received 21 October 2017
Revised 12 June 2018
Accepted 2 July 2018
Available online 21 July 2018

Keywords:
Deceptive review detection
Topic-sentiment joint probabilistic model
Latent dirichlet allocation
Gibbs sampling

ABSTRACT

In electronic commerce, online reviews play very important roles in customers' purchasing decisions. Unfortunately, malicious sellers often hire buyers to fabricate fake reviews to improve their reputation. In order to detect deceptive reviews and mine the topics and sentiments from the reviews, in this paper, we propose an *unsupervised topic-sentiment joint probabilistic* model (UTSJ) based on *Latent Dirichlet Allocation* (LDA) model. This model first employs *Gibbs sampling* algorithm to approximate parameters of maximum likelihood function offline and obtain topic-sentiment joint probabilistic distribution vector for each review. Secondly, a Random Forest classifier and a SVM (*Support Vector Machine*) classifier are trained offline, respectively. Experimental results on real-life datasets show that our proposed model is better than baseline models such as *n-grams, character n-grams in token*, POS (*part-of-speech*), LDA, and JST (*Joint Sentiment/Topic*). Moreover, our UTSJ model outperforms or performs similarly to benchmark models in detecting deceptive reviews over balanced dataset and unbalanced dataset in different domains. Particularly, our UTSJ model is good at dealing with real-life unbalanced big data, which makes it very suitable for being applied in e-commerce environment.

1. Introduction

With the popularity of intelligent mobile devices and the prosperity of the logistics industry, electronic commerce has recently become the main platform for shopping. However, information asymmetry is an open problem in the virtual market because sellers know the real quality of products while buyers may know little about the products or the sellers' reputation. Therefore, inexperienced buyers tend to refer to historical reviews which can guide their purchase decisions. However, some malicious sellers attempt to fabricate fake or deceptive reviews to improve their reputation. These reviews will not only mislead consumers' decisions but also disrupt markets with unfair competition. Due to the big volume of online reviews, it is a challenge to detect whether a review is fake or true in real time. Thus, automatic filtering of deceptive reviews and discovering aspects and corresponding sentiments of reviews deserve further studies.

In recent years, to address these problems, many researchers have proposed different offline learning models to evaluate the authenticity of customers' reviews. At first, the detection of deceptive review is regarded as a stylistic classification task from a traditional text classification viewpoint. Some researchers proposed the *n-grams* model (Choi & Kim, 2015; Elberrichi, 2006; Zhang & Wu, 2016) and the *character n-grams model* (Hernández Fusilier, Montes-y-Gómez, Rosso, & Guzmán Cabrera, 2015a) to mine text features. Then, some researchers proposed a *character n-grams in tokens* model (Cagnina & Rosso, 2015, Cagnina & Rosso, 2017), which is different from the traditional NLP feature *character n-grams* model, in consideration of tokens for feature extraction. Subsequently, some researchers used shallow syntactic features (Ghosh, Tonelli, & Johansson, 2013; Liu, Wei, Liu, & Fu, 2015) and deep syntactic features (Feng, Banerjee, & Choi, 2012; Li, Dan, & Hovy, 2015) in designing these classifiers. In particular, researchers (Ott, Choi, Cardie, & Hancock, 2011) innovatively proposed a model that integrated psychology and computational linguistics. Existing approaches mainly focus on traditional

* Corresponding author at: College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao, P.R. China.
E-mail addresses: 2281514572@qq.com (L.-y. Dong), jane_ji2003@aliyun.com, jsjsuzie@sina.com (S.-j. Ji), 601041109@qq.com (Q. Zhang), dicksonchiu@ieee.org (DicksonK.W. Chiu).

https://doi.org/10.1016/j.eswa.2018.07.005
0957-4174/

discrete features, which are based on linguistic and psychological cues. However, these methods failed to mine the semantics of a document from the discourse perspective. Therefore, they cannot adequately obtain implicit information from the reviews. To mine implicit information from reviews, some researchers (Li, Cardie, & Li, 2013; Li, Xie, Sun, & Bai, 2011) utilized the *Latent Dirichlet Allocation* (LDA) model to discover topic information from documents. However, in the reviews, reviewers not only discussed topics (i.e., aspects) of products but also expresses their feelings about these aspects of product.

As we all know, LDA model has good extensibility and sound mathematic basis. In order to improve performance of deceptive review detection, we extend the LDA model and apply it in detecting deceptive reviews. In this paper, we hypothesize that "*reviewers' sentiments are dependent upon topics to a great extent*" under the inspiration of human habit of writing (Li, Huang, & Zhu, 2010; Mei, Ling, Wondra, Su, & Zhai, 2007). For example, *human often choose topics such as location, cleanness, quietness, and so on, and then express their sentiment such as positive, negative, or neutral about these topics*. Based on this hypothesis, we present an *unsupervised topic-sentiment joint probabilistic model* (UTSJ) to mine topic-sentiment joint probabilistic distribution and further utilize this feature vector to distinguish truthful from deceptive.

In contrast to existing feature-based models, the innovation of this model is as follows. The UTSJ model first considers the topic (i.e., aspect) and then the sentiment of each review, which is consistent to the written expression habit of human and different to existing models especially *Joint Sentiment/Topic* (JST), which first considers the sentiment and then the topic of each review. The main difference between the JST and our UTSJ models lies in the computation principle of the sentiment joint topic distribution. The feature vector of the JST model considers distribution over topics for sentiment level, while our feature vector of UTSJ model given in this paper considers distribution over sentiment for topic level. Secondly, as demonstrated by our experiments, our UTSJ model outperforms the baseline feature-based models in *Precision, Recall,* and *F1-score*, especially in detecting real-life unbalanced deceptive reviews. Thirdly, in contrast to neural network-based representation learning algorithms, our UTSJ model can explain the detection (i.e., binary classification) results very well according to the topic-sentiment joint probability distribution.

The rest of this paper is organized as follows. Section 2 reviews the literatures in the area of detecting deceptive reviews. In Section 3, we introduce the USTJ model and the parameter estimation method. Section 4 illustrates the settings of experiments and analyzes the experimental results. We end this paper with conclusions and future prospects in Section 5.

2. Related work

Existing deceptive review detection models can be generally divided into two categories. One is the conventional feature-based models that classify review texts into deceptive ones or authentic ones. The other is neural network-based representation learning algorithms, which learn the word, sentence, or document level representation of review and detect spam opinions. We review these two categories briefly as follows.

(1) Feature-based models

In the area of feature-based deceptive review detection, a premier work (Jindal & Liu, 2008) analyzed the characteristics of spam activities and presented some novel detection techniques. Without labeled reviews, they tried to train models based on the linguistic, behavioral, and relationship among reviews, reviewers, and products. Ott et al. (2011) collected the first large-scale crowdsourcing dataset (called gold-standard opinion spam dataset) for deceptive opinion spam detection. This dataset includes truthful and deceptive reviews from the 20 most popular hotels on Trip Advisor. To classify the opinions into spam or non-spam, they further proposed an algorithm integrating concepts from psychology and computational linguistics fields and compared their algorithm with the n-grams model and the part-of-speech model. They used Linguistic Inquiry and Word Count (LIWC) software (Boyd & Pennebaker, 2016; Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007) to derive psychological and linguistic features. Due to the scarcity of deceptive reviews, Hernández, Fusilier, Montes-Y-Gómez, Rosso, and Guzmán Cabrera (2015b) employed *PU-learning* (which is a semi-supervised learning technique) in the detection of deceptive reviews, and adopted *character n-grams* as features to capture lexical content as well as stylistic information. They showed that *character n-grams* are better than word *n-grams* in the detection of opinion spam. Besides, Cagnina and Rosso (2017) proposed a *character n-grams in token* method to avoid the need of feature dimension reduction, and experimental results with intro and cross-domain cases showed that it is possible to obtain comparatively good results with a small amount of features.

The above researchers mainly focused on shallow lexicon-syntactic patterns. Alternatively, Feng et al. (2012) investigated a syntactic stylometry pattern for detecting deceptive reviews from an unconventional point by extracting features from *Context Free Grammar* (CFG) parse trees. Experimental results found that the existence of statistic signals hidden in deep syntax is helpful in distinguishing truthful reviews from deceptive ones. However, this method still cannot obtain the semantics of sentences. Moreover, as these methods ignored the intention under the literal features, it is still limited to traditional linguistic models. To address this problem, Mukherjee, Dutta, and Weikum (2016) designed a model that integrate linguistic *n-grams* features of review text with behavioral features of reviewers. They showed that this model outperformed those models only considering linguistic *n-grams* features on the real-life dataset of Yelp hotel and restaurant.

Topic modeling is a computational technique aimed at determining how different words can appear together to form a larger shared meaning (Pennebaker, Facchin, & Margola, 2010). It assumes that documents are under a distribution of some topics and each topic is a probabilistic distribution of terms (i.e., words). Without labeling opinions, this model can automatically analyze topic features of texts. Probabilistic topic models have experienced several stages of development such as *Latent Semantic Analysis* (LSA) (Deerwester, 1988), *Probabilistic Latent Semantic Analysis* (pLSA) (Hofmann, 1999), *Latent Dirichlet Allocation* (LDA) (Blei, Ng, & Jordan, 2003; Li et al., 2011), and *Hierarchical Dirichlet Process* (HDP) (Bartcus, Chamroukhi, & Glotin, 2015; Johnson & Willsky, 2012). As LDA model is an excellent mathematical model with high scalability, many researchers in text mining prefer to extend and apply it in sentiment analysis. For example, Titov and Mcdonald (2008) proposed a *Multi-Grain* LDA topic model, which not only extracted ratable aspects but also clustered them into coherent topics. The Multi-Grain LDA model classified topics into global topics and local topics. So, a word was sampled either from global topics or from the mixture of local topics specific for the local context of the word. This model is good at modeling topics or ratable aspects of online reviews (e.g., appropriate ratable topics for Italian restaurant could be *pizza* and *pasta*, whereas for Japanese restaurants they are probably *sushi* and *noodles*). But, this model still cannot learn the whole topic distribution for a review. In another words, no feature vector can adequately represent the topic distribution of a review for completing the classification task.

Lin and He (2009) proposed *Joint Sentiment/Topic model* (JST) model to mine the pairs of sentiment and topic. Similarly, Jo and Oh (2011) gave an *Aspect and Sentiment Unification Model* (ASUM) model to discover pairs of (aspect, sentiment). Lin, He, Everson, and Ruger (2012) further proposed a reparameterized version of

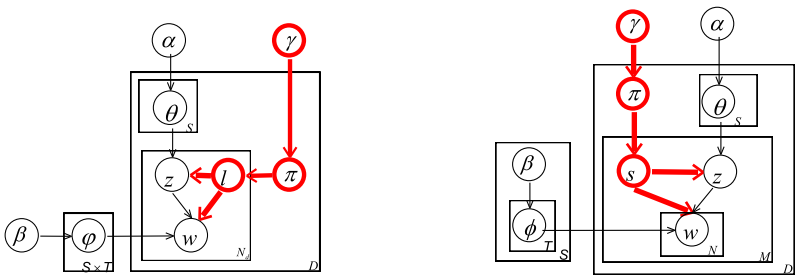


Fig. 1. Graphical representation of two improved LDA models.

the JST model called Reverse-JST, which is a weakly-supervised joint sentiment-topic model as it added sentiment prior information. Both the JST and Reverse-JST models were designed for general sentiment classification. The document sentiment is classified based on $p(l|d)$, the probability of a sentiment label given a certain document. Mukherjee et al. (2016) used the JST model to cast the review text into a number of informative facets and integrated this information with item ratings and timestamps in the detection of deceptive reviews. Although they used sentiment and aspect information derived from the JST model to detect deceptive reviews, their feature vector was different to ours in this paper in that, their topic-sentiment vector of each review was obtained by computing topic-sentiment-word distribution rather than the sentiment joint topic distribution learned from the JST model. The contribution of our model lies in that it can learn from the latent sentiment-topic distribution. So, we can directly use the learned sentiment-topic distribution instead of just computing the intermediate result.

Fig. 1(a) and Fig. 1(b) depict the graphical representation of JST model and ASUM model, respectively. The red thick part of the diagrams represents the extension of LDA model that incorporates sentiment. We can see that both models assume topic relies on sentiment. In addition, the ASUM model assumes that words in a single sentence must come from the same language model (Jo & Oh, 2011). Moreover, these two models focused on sentiment classification while neglecting text classification. To extract topic and corresponding sentiment features from reviews and apply it in detecting deceptive reviews, this paper proposes an *unsupervised topic-sentiment joint probabilistic model*. Section 3 will explain the main idea of this model in detail.

(2) Neural network-based algorithms

Neural network-based algorithms are popular in recent years. *Convolutional Neural Networks* (CNN) have been widely used for semantic composition (Johnson & Zhang, 2014; Kalchbrenner, Grefenstette, & Blunsom, 2014) and automatically capturing n-grams information. Sequential models such as *Recurrent Neural Network* (RNN) or *Long Short-Term Memory* (LUTSJ) have also been used for recurrent semantic composition (Li, Luong, & Dan, 2015; Tang, Qin, & Liu, 2015). The granularity of these representation learning algorithms also range from word (Sun, Du, & Tian, 2016) to sentence (Li, Qin, Ren, & Liu, 2017) and document (Ren & Ji, 2017) levels.

For example, the opinion spam detection system proposed by Ren and Ji (2017) is composed of three stages. In the first stage, a convolutional neural network is used to generate sentence representations from word representations. Then a bi-directional gated recurrent neural network is used to construct a document representation from the sentence vectors by modeling their semantic and discourse relations. Finally, the document representation is used as features to identify deceptive opinion spam. Such automatically induced dense document representation is compared with traditional manually-designed features on simulated datasets The detection model presented by Li et al. (2017) is a *Sentence Weighted Neural Network* (SWNN) model to learn the document-level representation of the review and detect spam reviews. They argue that learning the representation of the document can capture the global feature and take word order and sentence order into consideration.

Neural network-based algorithms can save researchers' time in analyzing the features of data. However, they lack the interpretation of the results. In contrast, the results obtained from the feature-based models have the properties of good interpretation and high computational efficiency. Moreover, the byproducts (e.g., topics and corresponding sentiments) of the detection is helpful in improving the quality of products and services. In addition, except for Mukherjee, Venkataraman, Liu, and Glance (2013) has tested and evaluated performance on the real-life Yelp dataset, all the above-mentioned models and algorithms have only been verified with simulated data, i.e., crowd-sourced fake reviews generated using *Amazon Mechanical Turk* (AMT) or by experts. It should be noted that the simulated data are balanced ones, i.e., the numbers of authentic and deceptive reviews in train dataset and test dataset are equal or nearly equal. However, in real-life, most of reviewers are honest and deceptive reviews is much less than authentic ones. For example, according to Luca and Zervas (2016), appropriately 16% of Yelp restaurant reviews are fraudulent. So, the algorithms that are good at classifying balanced datasets may not be suitable for unbalanced ones.

3. Unsupervised topic-sentiment joint probabilistic model

As we all know, customer comments on products in order to express their feelings of experience. This kind of written expression is to express one's sentiment and thought on some special topics within a certain period of time (Li et al., 2010; Mei et al., 2007). Inspired by this kind of human expression habit, we propose following hypothesis.

Hypothesis 1. In each review, the expressed sentiment is dependent on the specific topic.

For example, in a sample review like "*For the price and taste, this is a great restaurant*," two facets (i.e., price and taste) are presented along with the sentiment (i.e., great) under the topic of restaurant. However, the facet of taste is not usually presented with hotels or museums. According to this phenomenon, we know that those people who write review on item usually concern about some aspects of topic and express their sentiment about this aspect. This hypothesis is reasonable. Therefore, we propose that the joint probability distribution of sentiment and topic is more appropriate to be adopted as features in designing classifier.

In this section, we present an unsupervised topic-sentiment joint probabilistic model (UTSJ). Different to existing LDA-related models (Blei et al., 2003; Jo & Oh, 2011; Lin & He, 2009), our UTSJ model classify reviews according to the extracted topic features as well as corresponding sentiment features. This section first

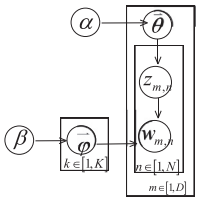


Fig. 2. Probabilistic graph model for LDA.

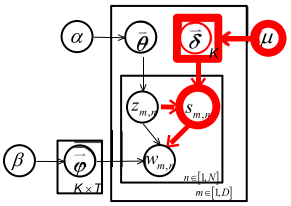


Fig. 3. Probabilistic graph model for UTSJ.

reviews the basic LDA model. Based on this model, we illustrate our topic-sentiment joint probabilistic model (UTSJ) proposed in this paper. Then, we explain the Gibbs sampling algorithm (Suess & Trumbo, 2010), which is used to efficiently implement the UTSJ model and to acquire the probabilistic distribution between sentiment and topics, as well as topics and words.

3.1. Basic LDA model

The LDA model is a generative model with three levels, namely, word, topic, and document. Definition 1 defines the LDA model formally.

Definition 1. A LDA model is a 6-tuple, LDA$=(\alpha,\ \beta,\ \vec{\theta},\ \vec{\varphi},\ z_{m,n},\ w_{m,n})$,

where

- α,β are hyper parameters that reflect the relative strength among implicit topics of document set, and the term probabilistic distribution about topic, respectively.
- $\vec{\theta}$ is a K-dimensional Dirichlet random variable, which is the probabilistic distribution of document-topic matrix.
- $\vec{\varphi}$ is a N-dimensional Dirichlet random variable, which is the probabilistic distribution of topic-word matrix.
- $z_{m,n}$ is the topic to which the n word in document m belongs.
- $w_{m,n}$ is the basic unit of discrete data that are defined to be an item indexed by n in document m.

Fig. 2 presents a typical directed probabilistic graph of a LDA model. The arc from a to $\vec{\theta}$ represents the process of generating a topic distribution matrix from a $Dirichlet(\alpha)$ distribution function for document d_m. The arc from $\vec{\theta}$ to $z_{m,n}$ describes the selection of a specific topic for word $w_{m,n}$ in document d_m. Given the topic $z_{m,n}$, the LDA model randomly selects a term through multinomial distribution of terms that is described by the arc from ϕ to$w_{m,n}$.

The implementation process of probabilistic model includes two steps:

(1) $\alpha \rightarrow \vec{\theta} \rightarrow z_{m,n}$: this step first generates $\vec{\theta}$ from a $Dirichlet(\alpha)$ distribution function and then samples $z_{m,n}$ from a $Multinomial(\vec{\theta})$ distribution function.
(2) $\beta \rightarrow \vec{\varphi} \rightarrow w_{m,n}$: this step intends to generate $\vec{\varphi}$ from a $Dirichlet(\beta)$ distribution function and then sample $w_{m,n}$ from a $Multinomial\vec{\theta}$ distribution function by given a specific$z_{m,n}$.

3.2. The UTSJ model

Though the LDA model can learn which aspects (or topics) reviewers concern, it cannot obtain the sentiment about the relative topics. Imaging a scenario when you want to write a review about a hotel, you may decide to describe several aspects of this hotel such as environment, condition, location, cleaness, and so on. Then, you will express different sentiments (positive or negative) for each aspect. For example, you may be satisfied with cleanliness but disappointed with the noisy environment of the hotel. In order to obtain the topic-sentiment joint probabilistic distribution of reviews, we improve the typical LDA model by adding sentiment levels to it. The proposed UTSJ model is a generative model with four levels, i.e., word, sentiment, topic, and document. Definition 2 define the UTSJ model formally.

Definition 2. A UTSJ model is a 9-tuple UTSJ$=(\alpha,\ \beta,\ \mu,\vec{\theta},\ \vec{\delta},\ \vec{\varphi},\ z_{m,n},\ s_{m,n},\ w_{m,n})$,

where

- α, β are hyper parameters that reflect the relative strength among implicit topics of document set, and the term probabilistic allocation about topic, respectively.
- μ is a hyper parameter that reflects the sentiment probabilistic distribution over topic.
- $\vec{\theta}$ is a K-dimensional Dirichlet random variable, which is the joint distribution of topic matrix.
- $\vec{\delta}$ is a T-dimensional Dirichlet random variable, which is the topic-sentiment joint distribution.
- $\vec{\varphi}$ is a N-dimensional Dirichlet random variable, which is the joint distribution of word matrix.
- $z_{m,n}$ is the topic that the n word in document m belongs to.
- $s_{m,n}$ is the sentiment that the n word in document m belongs to.
- $w_{m,n}$ is the basic unit of discrete data, which is defined to be an item indexed by n in document m.

Fig. 3 presents a directed probabilistic graph of UTSJ model. The red thick part of the graph represents the extension part of basic LDA model. In Fig. 3, the arc from α to $\vec{\theta}$ describes the process of selecting a certain topic $z_{m,n}$ through topic multinomial distribution for every word $w_{m,n}$ in each document d_m. The arc from $\vec{\delta}$ to $s_{m,n}$ describes the process of selecting a sentiment for every word $w_{m,n}$through the sentiment multinomial distribution given the topic$z_{m,n}$. The arc from $\vec{\varphi}$ to $w_{m,n}$ defines the selection of a term through term multinomial distributions given the topic $z_{m,n}$ and sentiments$_{m,n}$. The main difference of Fig. 3 from Fig. 1(a) is that the model in Fig. 3 obtains the topic first and then user's sentiment over the topic, while the models in Fig. 1(a) obtain the sentiment first and then the topic.

The detailed generating process of UTSJ is as follows:

(1) For each topic $k \in \{1, 2, \cdots K\}$, draw a K-dimension random variable $\vec{\theta} \sim Dirichlet(\alpha)$;
(2) For each sentiment $t \in \{1, 2, \cdots T\}$, draw a T-dimension random variable $\vec{\delta} \sim Dirichlet(\mu)$;
(3) For each word $w_{m,n}$ in document m:
   i. Select topic $z_{m,n} \sim Multinomial(\vec{\theta})$;
   ii. According to $z_{m,n}$, select sentiment $s_{m,n}$, where $s_{m,n} \sim Multinomial(\vec{\delta})$;
   iii. Given $z_{m,n}$and $s_{m,n}$, generate word $w_{m,n}$, where $w_{m,n} \sim Multinomial(\vec{\varphi})$;

The parameters$\vec{\theta}$, $\vec{\delta}$, and $\vec{\varphi}$ in above steps should be estimated according to Gibbs sampling algorithm. The derivation process and the steps of Gibbs sampling algorithm are given in Section 3.3.

3.3. Parameter estimation

In the UTSJ model, the Maximum Likelihood Estimation (MLE) process is used to estimate parameters $\vec{\theta}$, $\vec{\delta}$, and $\vec{\varphi}$. These estimation process can be realized by following steps. (1) $\alpha \to \vec{\theta} \to z_{m,n}$. It represents two structures, i.e. $\vec{\theta} \sim Dirichlet(\alpha)$ and $z_{m,n} \sim Multinomial(\vec{\theta})$. Moreover, the generating process of topic mentioned above (i.e., the flow of $\alpha \to \vec{\theta} \to z_{m,n}$) for different document is independent. Therefore, for each document, the probability of topics is generated according to Eq. (1).

$$p\left(\vec{z}|\alpha\right) = \prod_{k=1}^{K} p\left(\vec{z_m}|\alpha\right) = \prod_{k=1}^{K} \frac{\triangle\left(\vec{n_{mk}} + \alpha\right)}{\triangle\left(\alpha\right)} \tag{1}$$

where,$\overrightarrow{n_{mk}} = (N_m^{(1)}, N_m^{(2)}, \cdots, N_m^{(K)})$, $N_m^{(k)}$ represents the number of words that are generated by topic k in document m, $z_{m,n}$ represents the topic that every word in each document belongs to. $\Delta(\alpha)(\alpha = \{\alpha_1, \alpha_2, \cdots \alpha_n\})$, which is the normalizing factor of $Dirichlet(\alpha)$ distribution function. That is to say, $\triangle\alpha$ can be calculated by Eq. (2):

$$\triangle\,\alpha = \int \prod_{k=1}^{K} {p_k}^{\alpha_k - 1} d\vec{p} \tag{2}$$

(2)$\mu \to \delta \to s_{m,n}$. This step includes two structures, i.e., $\mu \to \vec{\delta}$, $\vec{\delta}_k \to s_{m,n}$. $\mu \to \vec{\delta}$ corresponds to a Dirichlet structure, and $\vec{\delta}_k \to s_{m,n}$ corresponds to a multinomial distribution. Therefore, $\mu \to \vec{\delta} \to s_{m,n}$ is a Dirichlet-multinomial conjugated structure. As we assume that the generating process of sentiment is independent of the topics, once given the generating probability of the topics, the probability of sentiment can be calculated according to Eq. (3):

$$p\left(\vec{s}\,|\,\vec{z}\,,\mu\right) = \prod_{t=1}^{T} p\left(\vec{s_t}\,|\vec{z_k},\mu\right) = \prod_{t=1}^{T} \frac{\triangle\left(\vec{n_{mkt}} + \mu\right)}{\triangle\left(\mu\right)} \tag{3}$$

where, $\vec{n_{mkt}} = (N_{m,k}^{(1)}, N_{m,k}^{(2)}, \cdots, N_{m,k}^{(t)})$, $N_{m,k}^{(t)}$ represents the number of words that are associated with the topic k and sentiment t in document m.

(3)$\beta \to \varphi \to w_{m,n}$. It has two structures, namely,$\overrightarrow{w_{m,n}} \sim Multinomial(\vec{\varphi})$, $\vec{\varphi} \sim Dirichlet(\beta)$. As we assume that the generation of words is mutual independent, the probability for words in the corpus could be calculated according Eq. (4).

$$p\left(\vec{w}|\vec{z},\vec{s},\beta\right) = \prod_{w=1}^{V} p\left(\vec{z_t}|\vec{z_k},\vec{s},\beta\right) = \prod_{w=1}^{V} \frac{\triangle\left(\vec{n_{ktw}} + \beta\right)}{\triangle\left(\beta\right)} \tag{4}$$

where,$\vec{n_{ktw}} = (N_{k,t}^{(1)}, N_{k,t}^{(2)}, \cdots, N_{k,t}^{(V)})$, $N_{k,t}^{(V)}$ represents the number of words that are allocated to the topic k, and sentiment t. $w_{m,n}$ represents the word in the n^{th} position of document m.

Comprehensive consideration of Eqs. (1) and (3) arrives at the joint probability of distribution of latent variables as defined in the following Eq. (5).

$$\begin{aligned} &p(w,z,s|\alpha,\mu,\beta) = p(w|z,s,\beta)p(s|z,\mu)p(z|\alpha) \\ &= \prod_{k=1}^{K} \frac{\triangle\left(\vec{n_{mk}} + \alpha\right)}{\triangle\left(\alpha\right)} \prod_{t=1}^{T} \frac{\triangle\left(\vec{n_{mkt}} + \mu\right)}{\triangle\left(\mu\right)} \prod_{w=1}^{V} \frac{\triangle\left(\vec{n_{ktw}} + \beta\right)}{\triangle\left(\beta\right)} \end{aligned} \tag{5}$$

Based on Eq. (5), we can be derive Eq. (6) by Gibbs sampling. $p(z_i = k,\ s_i = t|z_{-i},\ s_{-i},\ w,\ \alpha,\ \beta) \propto p(z_i = k,\ s_i = t,\ w_i = v|z_{-i},\ s_{-i},\ w_{-i},\ \alpha,\ \beta)$

$$= \frac{\{\vec{n_{mk}}\}_{-i} + \alpha}{\sum_K \left(\{\vec{n_{mk}}\}_{-i} + \alpha\right)} \frac{\vec{n_{mkt}} + \mu}{\sum_T \left(\{\vec{n_{mkt}}\}_{-i} + \mu\right)} \frac{\vec{n_{ktw}} + \beta}{\sum_V \left(\{\vec{n_{ktw}}\}_{-i} + \beta\right)} \tag{6}$$

The basic process of Gibbs sampling algorithm Suess &Trumbo, 2010) adopted to learn the above parameters is given in Algorithm 1. Inside the three-layer loop structure, this algorithm sequentially samples topics and sentiment according to Eqs. (1) and ((3), respectively. After updating the statistic variables, this algorithm calculates the parameters θ,δ, and ϕ according to Eqs. (7)-(9). The computational time complexity of this algorithm is O (maxIter* documentNo*wordsNo), where maxIter denote maximal iterations of Gibbs sampling, documentNo represents the number of document, wordsNo represents the length of document.

Taking the expected value of parameters θ,δ, and ϕ, which exist in the posterior distribution of Eq. (6) as estimated value of them, we can get following equations.

$$\theta_m = \frac{\{\vec{n_{mk}}\}_{-i} + \alpha}{\sum_K \left(\{\vec{n_{mk}}\}_{-i} + \alpha\right)} \tag{7}$$

$$\delta_k = \frac{\{\vec{n_{mkt}}\}_{-i} + \mu}{\sum_T \left(\{\vec{n_{mkt}}\}_{-i} + \mu\right)} \tag{8}$$

$$\varphi_{k\times t} = \frac{\{\vec{n_{ktw}}\}_{-i} + \beta}{\sum_V (\{\vec{n_{ktw}}\}_{-i} + \beta)} \tag{9}$$

where,$\{\vec{n_{mk}}\}_{-i}$, $\{\vec{n_{mkt}}\}_{-i}$, $\{\vec{n_{ktw}}\}_{-i}$are sufficient statistics as explained in Table 1. Further, θ_m, δ_k, and $\phi_{k\times t}$ are the three Dirichlet posterior distribution under the Bayesian framework, i.e., θ_m is the document-topic probabilistic distribution,δ_k is the topic-sentiment joint probabilistic distribution, and $\phi_{k\times t}$ is the topic-sentiment-word probabilistic distribution.

4. Experiments and results

To evaluate the performance of our UTSJ model given in this paper and to further compare it with typical feature-based ones in the field of deceptive review detection, such as *unigram* (Ott et al., 2011), *character n-grams in token* (Cagnina & Rosso, 2017), POS (Ott et al., 2011), LDA (Li et al., 2013), and the JST model (Lin & He, 2009). Note that the JST and Reverse-JST models are most similar to our model. However, the reverse-JST model are not chosen as benchmark because it is a weak supervised model, while the above models including ours are unsupervised ones. We design and implement three sets of experiments. In the first set of experiments, we compare the differences among the LDA, JST, and UTSJ models in perplexity. These models are generative probabilistic model. As there are two parameters (i.e., number of Gibbs sampling iterations and number of topics) in these three models, in the first set of experiments, we focus on observing the variation characteristics of perplexity with the increase of the number of Gibbs sampling iterations and the number of topics. The second set of experiments is to evaluate the classification performance over balanced and unbalanced restraurant dataset under criteria such as *Precision, Recall*, and *F1-Score*. The aim of the third set of experiments is to verify the classification performance of the compared models over the Yelp dataset from different domain (i.e., hotel), which also verifies the adapbility of these models.

4.1. Data description and experimental settings

The empirical data are labeled English reviews from yelp.com. Table 2 shows the statistic characteristics of this dataset. The

Algorithm 1 Gibbs sampling Algorithm.

Input: 1) hyper-parameters α, μ, β; 2) maximal iterations (denoted as maxIter) of Gibbs sampling;

Output: matrixes θ, δ, φ

Process:

```
Randomly initialize matrix θ, δ, φ
For i = 1 to maxIter
    For all document
        For all words in document
            Sampling new topic according to Equation (1)
            Sampling new sentiment according to Equation (3)
            Update statistics: {n_mk}_-i, {n_mkt}_-i, {n_ktw}_-i
            Calculate θ according to Equation (7)
            Calculate δ according to Equation (8)
            Calculate φ according to Equation (9)
        End for
    End for
End for
```

Table 1
Definition of characters in Gibbs sampling.

Character	Definition
$z_i = k$	The i-th word belong to topic z
$s_i = t$	The i-th word belong to sentiment s
$w_i = v$	the$i$-th word is word $v$in dictionary
z_{-i}	except for the$i$-th word, the remaining words in the document and the corresponding topic relationship
s_{-i}	except for the$i$-th word, the remaining words in the document and the corresponding sentiment relationship
w_{-i}	except for the$i$-th word, the remaining words in the document and the corresponding relationship with dictionary
$\{\overrightarrow{n_{mk}}\}_{-i}$	except for the$i$-th word, the number of words generated by the$k$-th topic in document$m$
$\{\overrightarrow{n_{mkt}}\}_{-i}$	except for the$i$-th word, the number of words generated by the$k$-th topic and$t$-th sentiment in document m
$\{\overrightarrow{n_{ktw}}\}_{-i}$	except for the$i$-th word, the number of word$v$ generated by the$k$-th topic and$t$-th sentiment

Table 2
Dataset statistics for review classification.

Dataset	Deceptive Reviews	Authentic Reviews	Deceptive%	Total reviews	used in experiments
HotelND	780	5078	13.3	5858	
RestaurantND	8303	58,716	12.4	60,719	
Hotel$^{\#}$	780	1170	40	1950	3nd set
Restaurant$^{\#}$	8303	12,454	40	20,757	2nd set
Hotel*	780	780	50	1560	3nd set
Restaurant*	8303	8303	50	16,606	1nd set, 2nd set

datasets have labels about whether a review is deceptive or not. The deceptive reviews are filtered ones obtained by Yelp's spam filter, while truthful reviews are from Yelp's regular webpages. These reviews come from two domains, i.e., hotel and restaurant. In the hotel domain, there are 780 deceptive reviews and 5,078 truthful reviews. For restaurant, it includes 8,308 deceptive reviews and 58,716 truthful review. From Table 2, we can see that the class distribution of the Yelp dataset is skewed. Dataset with 'ND' denotes natural distribution. As it is well known that highly unbalanced data often produces poor models (Mukherjee et al., 2013), to build a good model for unbalanced data, we employ under-sampling (Drummond & Holte, 2003) to construct unbalanced datasets. Under-sampling is a common method employed to randomly select a subset of instances from the majority class and combine it with the minority class to form a balanced class distribution data for model building. For example, Mukherjee et al. (2013) used this naïve method in constructing unbalanced dataset over real-life Yelp dataset. In Table 2, the unbalanced dataset and balanced dataset generated with under-sampling are labeled with "#" and "*", respectively. To verify the adaptability and generalization performance of the models, this paper implements two set of experiments on data from two domains. The hotel dataset is used in the first and second set of experiments, while the restaurant dataset is used in the third set of experiments. Before implementing the three sets of experiments, we first pre-process the reviews by removing digits and punctuations. Then we separate words by blank space as well as get the part-of-speech of each word.

In the implementation of these three sets of experiments, all the classification tasks are implemented by using 5-fold *Cross Validation* (CV). The models *unigram, character n-grams in token (abbr. C-ngrams-token)*, POS, LDA, and JST, which have been reviewed in related work section, are selected as baseline models, because they are typical representatives of various feature-based detection methods. We adopt the Random Forest classifier for all the experiments. Especially, for the high-dimensional features such as *unigram* model and *character n-grams in token* model, we also

experiment with a SVM classifier. That is because SVM is appropriate to deal with high-dimensional features (Cortes & Vapnik, 1995).

In all experiments, the optimal values of the parameters in baseline models are adopted. For the models of *unigram, character n-grams in token*, and POS, all features are encoded with TF-IDF values (which is because it is a common method to preprocess such text). For the LDA model, we take topic probabilistic distribution (represented by $\bar{\theta}$ in Fig. 2) as the feature of a review, which is computed according to Eq. (7). For the JST and UTSJ models, we take sentiment-topic probabilistic distribution and topic-sentiment probabilistic distribution (respectively represented by $\bar{\theta}$ and $\bar{\delta}$) as the feature of a review. The feature vector of the UTSJ model is computed according to Eq. (8). Similar to the setting of Lin and He (2009), in the three sets of experiments, Dirichlet hyper-parameters are assigned with 0.1, 0.01, and 0.1, respectively. Besides, in implementation topic models, the number of topics is assigned with 5, 10, 15, and 20, respectively.

4.2. Evaluation criteria

Similar to Ott et al. (2011) and Feng et al. (2012), we choose *Precision, Recall*, and *F1-Score* as evaluation criteria, which is defined in Eqs. (10), (11), and (12), respectively. The larger these values, the better the classifier is.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{12}$$

where, TP (True Positive) refers to the number of positive tuples classified correctly as positive by the classifier; TN (True Negative) refers to the number of negative tuples classified correctly as negative by the classifier; FP (False Positive) refers to the number of negative tuples wrongly labeled as positive; and FN (False Negative) refers to the number of positive tuple wrongly labeled as negative. Reported values of *Precision, Recall*, and *F1-Score* are computed using a macro-average.

In particular, we computed the perplexity of a held-out test set to evaluate the models. The perplexity, which is used in language modeling, is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity score indicates a better generalization performance. Formally, for a test set of M documents, the perplexity is calculated as follows.

$$Perplexity(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{13}$$

4.3. Results and analysis

4.3.1. Results when changing the number of gibbs sampling iterations and the number of topics

Fig. 4(a) illustrate the results we get from the first set of experiments, which aims at finding differences among the LDA, JST, and UTSJ models in perplexity. The horizontal axis is the number of iterations, and the vertical axis represents the values of perplexity. As Fig. 4(a) shows, the curves of the LDA, JST, and UTSJ models are decreasing with the increase of iterations. As we know, a lower perplexity score indicates a better performance in general. Seen from Fig. 4(a), the perplexity of our UTSJ model is always smaller than that of the LDA and JST models. This demonstrates that our UTSJ model is superior to the other two models in generalizing performance. It should be noted that the speed of decrease is quick when the number of iteration is less than 40, but becomes slow afterwards. When the number of iterations increase to 500, the perplexity almost keep stable, which means that these models are basically converged. Therefore, we set iterations to 500 in the second and third set of experiments.

Similar to the characteristics that the number of Gibbs Sampling iterations influences perplexity, Fig. 4(b) shows a similar tendency of perplexity when the number of topics varies from 5 to 20, where the horizontal and vertical axes represent the number of topics and the value of perplexity, respectively. Generally speaking, the curves of perplexity tend to decrease as the number of topics increase. Similar to the previous evaluation criterion, the curve of our UTSJ model is always under that of the other two models, which means that our UTSJ model is superior to the other models. When the number of topics is less than 15, the decrease speed of perplexity is quick, but becomes slow after that. Therefore, these models can achieve almost their best performance when the topic is assigned with an appropriate value (e.g., 15).

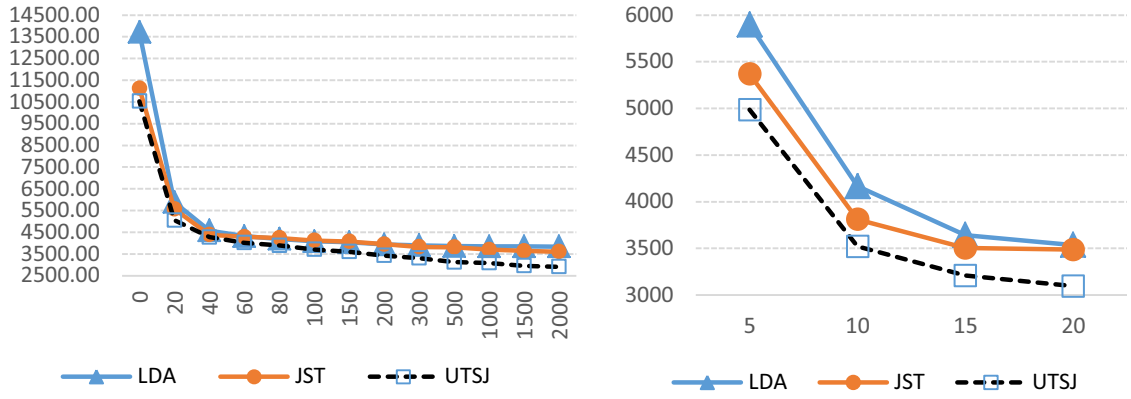
4.3.2. Results of comparison of the models over balanced and unbalanced datasets

In the second set of experiments, we compare the performance of our model with baseline models using balanced and unbalanced restaurant dataset (see Table 2), respectively. As the performance of the LDA, JST and UTSJ are influenced by the number of topics, it is necessary to compare the performance of these models under different number of topics. Figs. 6 and 7 show the results with balanced and unbalanced restaurant datasets, respectively, when the number of topics range from 5 to 20. The horizontal axis is the number of topics. In addition, the vertical axis represents the values of *Precision, Recall*, and *F1-Score* that are obtained under various number of topics. From the results of the first set of experiments (see Fig. 5), it can be seen that most topic models can achieve their best performance when the number of topics is set to 15. To further illustrate the performance comparison of these models in detail, we especially list the *Precision* (denoted as P), *Recall* (denoted as R), and *F1-score* (denoted as F) values in 3 when the number of topics is set to 15. All the values in following figures and tables are represented in percentage (%), which is commonly used in the literature.

Results over balanced restaurant dataset

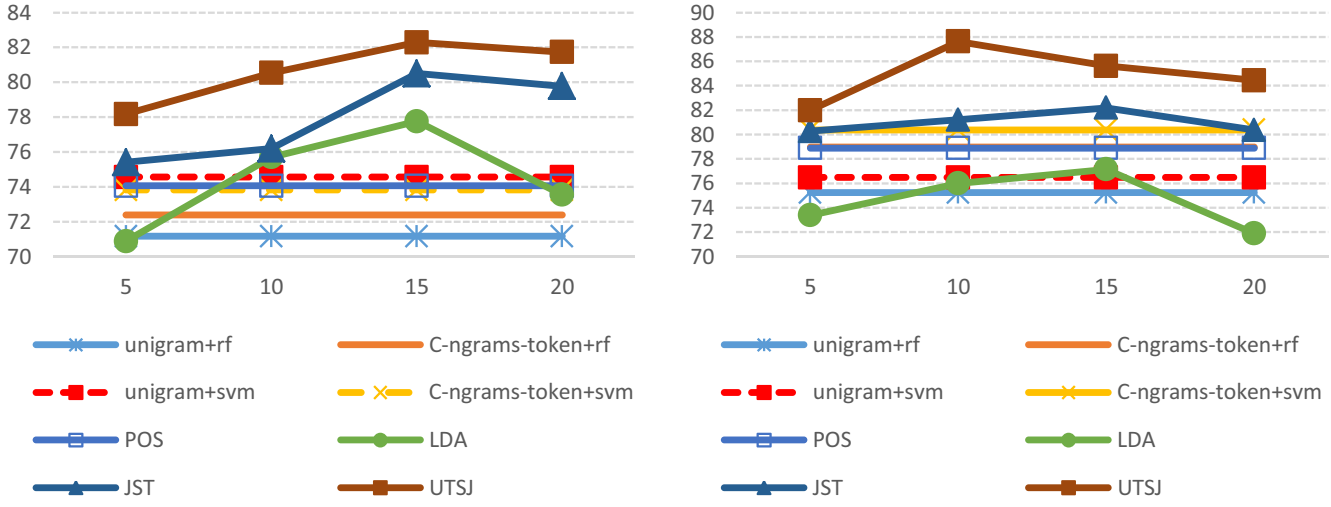
In Fig. 5, the *unigram, character n-grams in token*, and *POS* models are not influenced by number of topics. From Fig. 5 and Table 3, we can see that when using Random Forest classifier, *character n-grams in token* (72.39, 78.99, and 75.57) model performs better than *unigram* model (71.16, 75.23, and 73.14) in *Precision, Recall*, and *F1-Score*. Similar performance difference occurs when SVM classifier is used. This is because the *character n-grams in token* model performs better than *n-grams* model in distinguishing the writing style of the deceivers. In comparison, the *Precision, Recall*, and *F1-Score* attained by POS (74.06, 78.89, 76.39) is slightly larger than the corresponding ones of the *unigram* and *character n-grams in token* models, because the POS model utilizes shallow syntax feature that mines implicit information in review contexts.

Among the three topic-related models, the LDA model (77.76, 77.14, and 77.45) is significantly inferior to the JST model (80.5, 82.19, and 81.34) and the UTSJ model (82.29, 85.62, and 83.92). This is because the JST and UTSJ models consider both topic and sentiment information. Comparing the results in the three subgraphs of Fig. 5, we can see that the precision of these topic models is slightly improved when the number of topics increases from 5 to 15. When the number of topics reaches 20, the LDA model has a sharp decline, whereas the JST and UTSJ models are more stable. Among the three topic models, our UTSJ model can always achieve a higher performance than the other topic-based models whatever the number of topics is assigned with. Moreover, from Fig. 5, we

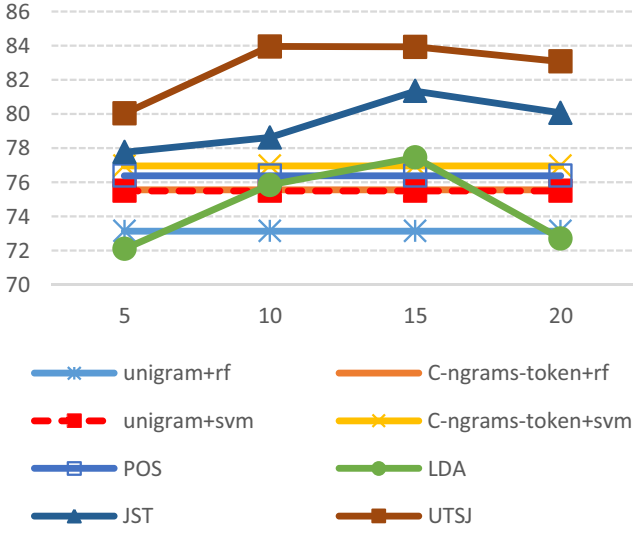


(a) Iterations of the Gibbs sampler **(b) Number of topics**

Fig. 4. Perplexity as a function of iterations of the Gibbs sampler and number of topics for LDA, JST and UTSJ.



(a) Precision (b) Recall



(c) F1

Fig. 5. Performance comparison of the models over balanced restaurant dataset.

Table 3
Performance comparison of the models with restaurant dataset when topic = 15.

Category	Model	*Balanced dataset*			*Unbalanced dataset*		
		P	*R*	*F*	*P*	*R*	*F*
topic unrelated models	Unigram + RF (Ott et al., 2011)	71.16	75.23	73.14	65.46	73.22	69.12
	C-ngrams-token + RF (Cagnina & Rosso, 2017)	72.39	78.99	75.57	67.41	69.63	68.5
	Unigram + SVM (Ott et al., 2011)	74.56	76.47	75.5	64.15	68.43	66.22
	C-ngrams-token + SVM (Cagnina & Rosso, 2017)	73.82	80.38	76.96	64.79	69.03	66.84
	POS (Ott et al., 2011)	74.06	78.89	76.39	72.89	80.27	76.4
topic related models	LDA (Li et al., 2013)	77.76	77.14	77.45	73.75	77.74	75.69
	JST (Lin & He, 2009)	80.5	82.19	81.34	72.24	74.19	73.2
	UTSJ (this paper)	**82.29**	**85.62**	**83.92**	**75.25**	**80.57**	**77.82**

can see that our UTSJ model outperforms all the topic-unrelated ones whatever the number of topics is assigned.

Conclusion 1. Our UTSJ model has the best performance among these models in detecting deceptive reviews when it is verified on balanced restaurant dataset.

Results over unbalanced restaurant dataset

From Fig. 6 and Table 3, we can also see that the POS model is superior to the *unigram* and *character n-grams in token* models as it utilizes part-of-speech feature of words. Moreover, when we adopt the Random Forest classifier, the *character n-grams in token* model (67.41, 69.63, and 68.5) gains slightly improvement in *Precision* over the *unigram* model (65.46, 73.22, 69.12) while being inferior in terms of *Recall* and *F1-Score*. When we apply SVM classifier, the *character n-grams in token* model (64.79, 69.03, and 66.84) is superior to the *unigram* model (64.15, 68.43, and 66.22), but still inferior to the results from the balanced dataset (see Table 3). We think that this phenomenon may be caused by the differences of the vocabulary size between the balance and imbalance datasets. When we construct an unbalanced dataset, the ratio of positive reviews (i.e., truthful reviews) is naturally much larger in the training set, and hence the vocabulary about positive reviews is also much larger. Therefore, for unbalanced datasets, it is more difficult to detect deceptive reviews.

Observing the trend of topics model when changing number of topics, we found that the JST model sometimes behaved worse than the LDA model (see Fig. 6). Besides, the LDA and JST models are more vulnerable than the UTSJ model. For example, when we set the number of topic to 20, all the values of the *Precision, Recall*, and *F1-score* obtained by using the LDA and JST models are correspondingly smaller than that with the number of topic set to 15. Instead, the performance of the UTSJ model is steadily increasing when the number of topics varies from 5 to 15. Especially, when the number of topics is set to 20, the *Precision, Recall*, and *F1-score* obtained by using the UTSJ model is almost similar to those with 15 topics. As we can see from the Fig. 6, when the number of topics is assigned with 5 and 10, POS is obviously higher than other models in *Precision, Recall* and *F1-Score*. When the number of topics is set to 15, the performance of topic models is improved. Meanwhile, our UTSJ model is best ones among other models. This indicates that number of topics is important parameter. Only if we set appropriate number of topics, the performance of topic models will be better.

Comparing the results listed in Table 3 from the horizontally view, it can be seen that the results we get using the unbalanced restaurant dataset is correspondingly smaller than those with the balanced dataset. For example, when we adopt *unigram* model together with Random Forest classifier, the results (65.46, 73.22, and 69.12) of the unbalanced dataset is lower than those (71.16, 75.23 and 73.14) of the balanced dataset. That is consistent with the common sense in nature language field that *it is more difficult to detect deceptive review under unbalanced dataset*. Carefully observing the results of unbalanced dataset listed in Table 3, we can see that our UTSJ model is the best among these models in detecting deceptive reviews with an unbalanced dataset, and we are significantly better than other models.

Conclusion 2. : When implementing our model on unbalanced dataset, it is the best topic related model with significant improvement.

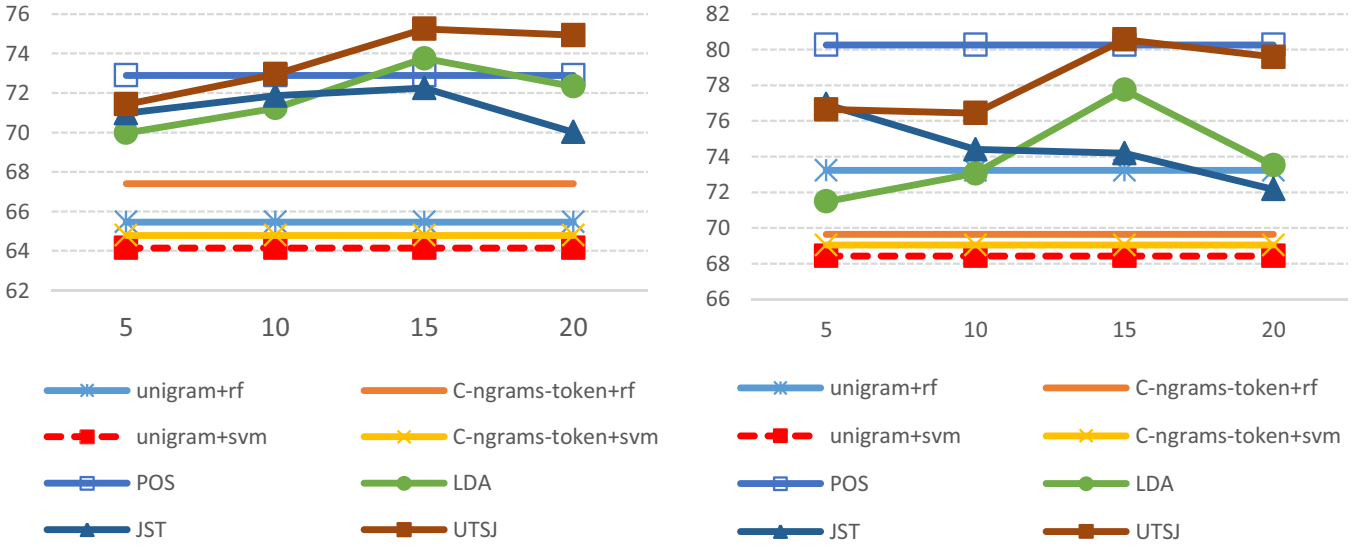
4.3.3. Results comparison of the models over different domain dataset

To further compare the adaptivity to different domains, we choose the hotel dataset listed in Table 2 as another domain to verify the performances with the third set of experiments. The setting of the third set of experiments is similar to that of the second set. Fig. 7 and 8 illustrate the results of our third set of experiments. The horizontal axis is the number of topics. In addition, the vertical axis represents the values of *Precision, Recall*, and *F1-Score* that are obtained under various number of topics. As the results of the first set of experiments (see Fig. 4) show that most topic-related models can achieve their best performance when the number of topics is assigned with 15, therefore, to illustrate the performance comparision of these models over hotel domain in detail, we specially list in Table 4 the *Precision* (denoted as P), *Recall* (denoted as R), and *F1-score* (denoted as F) values when the number of topics is set to 15. All the values in following figures and tables are represented in percentage (%), which is commonly used in the literature.

Results over balanced hotel dataset

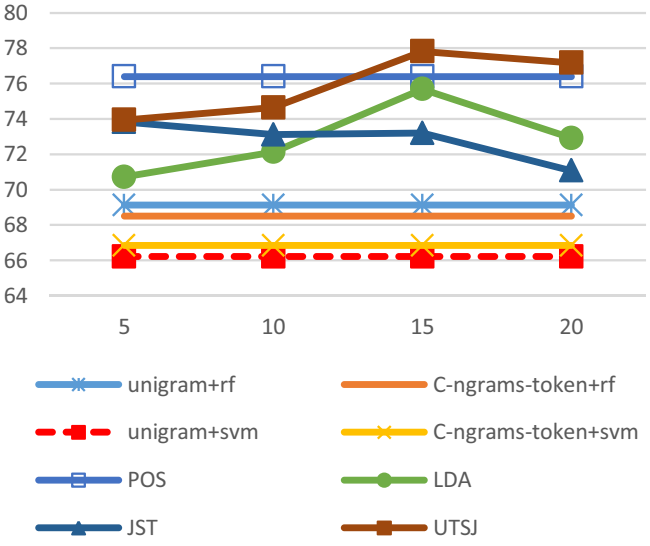
Firstly, we will analyze the performance of the topic-unrelated models. In comparison, the *Precision, Recall*, and *F1-Score* attained by the *character n-grams in token* model (71.3, 79.52, and 75.19) is slightly larger than the corresponding ones of the *unigram* model (70.42, 75.63 m and 72.93) when we apply Random Forest classifier. This is because the *character n-grams in token* model can not only keep the main characteristic of the standard *n-grams* model, but also can obtain smaller features than the *unigram* model. In comparison, the POS model is better than the *unigram* and *character n-grams in token* models, which indicates that simple genre identification approach do help in detecting deceptive reviews.

Comparing the results (76.34, 85.53, and 79.77) of the LDA model to that (75.92, 82.42, and 79.04) of the POS model, we can see that the LDA model is also superior to the POS model in *Precision, Recall*, and *F1-Score*. That is because the LDA model can capture hidden semantic information in reviews. As we can see, when the number of topic is as large as 20, the *Precision* of LDA model decrease sharply and the *Recall* increase somewhat (see Fig. 7). This indicates that the LDA model is greatly influenced by the number of topics. Compared to the results of the LDA model (76.34, 85.53, and 79.77), the JST model (83.5, 84.75, and 84.12), and the UTSJ model (87.15, 83.02 and 85.03), it can be seen that the models that not only consider semantic information but also sentiment polarities can achieve much better performance. Besides, our UTSJ model is the best among the compared models over the



(a) Precision

(b) Recall



(c) F1

Fig. 6. Performance comparison of models on unbalanced restaurant dataset.

Table 4
Performance comparison of the models with hotel dataset.

Category	Model	*Balanced dataset*			*Unbalanced dataset*		
		P	*R*	*F*	*P*	*R*	*F*
topic unrelated models	*Unigram* + RF (Ott et al., 2011)	70.42	75.63	72.93	73.61	75.2	74.39
	C-ngrams-token + RF (Cagnina & Rosso, 2017)	71.3	79.52	75.19	75.51	76.23	75.87
	Unigram + SVM (Ott et al., 2011)	72.94	73.43	73.18	71.29	74.68	72.95
	C-ngrams-token + SVM (Cagnina & Rosso, 2017)	73.86	72.15	72.99	74.56	78.94	76.69
	POS(Ott et al., 2011)	75.92	82.42	79.04	78.93	86.97	82.75
topic related models	LDA(Li et al., 2013)	76.34	83.53	79.77	78.91	86.83	82.68
	JST(Lin & He, 2009)	83.5	**84.75**	84.12	81.83	86.92	84.29
	UTSJ (this paper)	**87.15**	83.02	**85.03**	**83.62**	**87.27**	**85.41**

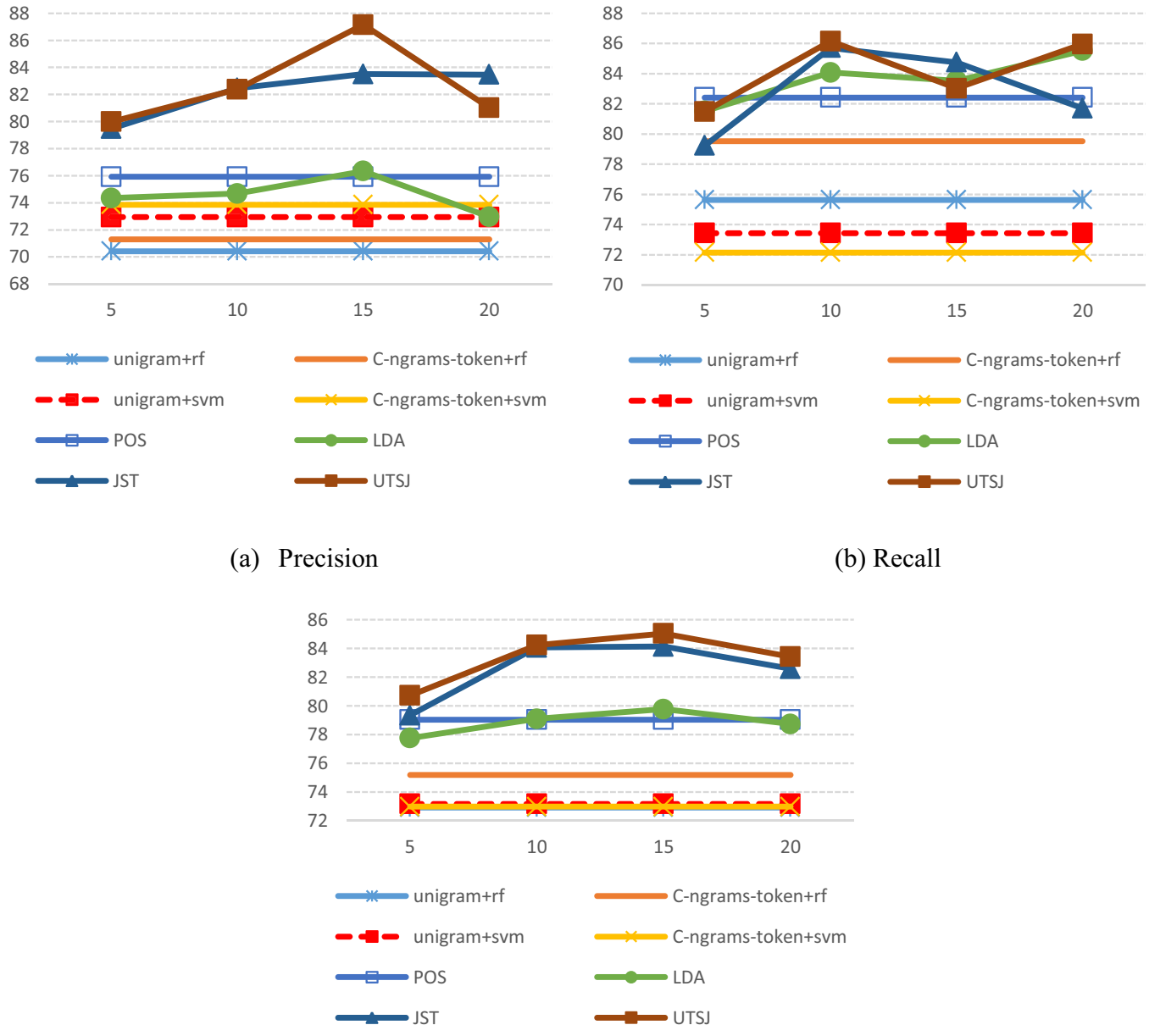


Fig. 7. Performance comparison of the models on balanced hotel dataset.

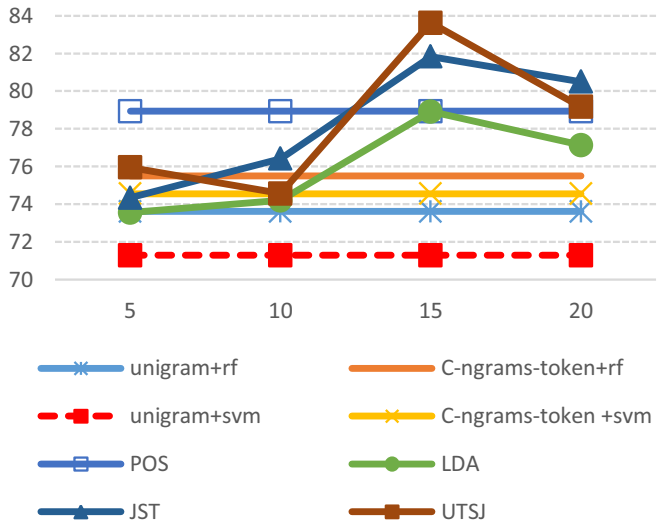
balanced hotel dataset. Therefore, we can conclude that simultaneous consideration of topic and sentiment distribution can further improve the performance of deceptive reviews detection.

Results over unbalanced hotel dataset

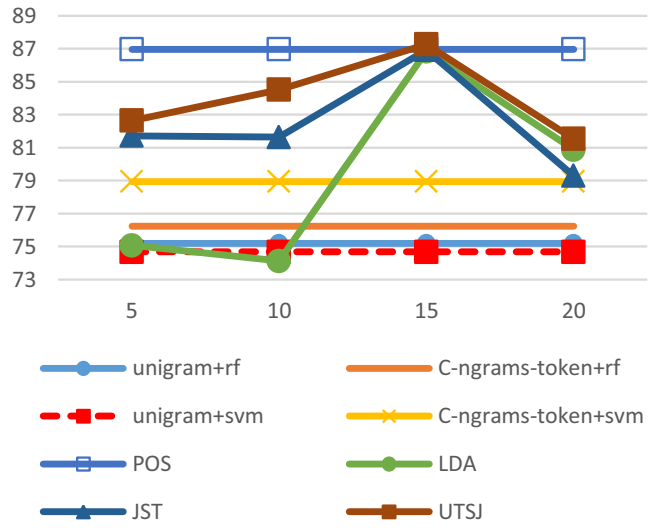
Fig. 8 show the performances of benchmark models on the unbalanced hotel dataset. The performance (i.e., *Precision, Recall*, and *F1*) of all models is correspondingly lower than those obtained over the balanced hotel dataset, because of the unbalanced proportion of fake samples as explained with the previous restaurant dataset. Similarly, the *Character n-grams in token* model is still better than the *unigram* model in capturing linguistic features of text. The POS model is also superior to the *unigram* model and *Character n-grams* models because the frequency of part-of-speech tags in a text is often dependent on the genre of the text.

When the number of topic is set to a value smaller than 15, the performance of topic-related models is not as good as the POS model, i.e., (78.93, 86.97, and 82.75) in Fig. 8. When the number of topics is increased to 15, the three topic-related models can achieve their optimal performances and outperform the POS model. Therefore, choosing an appropriate number of topic is crucial to get good results in detecting deceptive reviews. From the results listed in each "*Unbalanced dataset*" column of Table 4, we can see that our UTSJ model gains the largest *Precision* (83.62), *Recall* (87.27), *F1-Score* (85.41) simultaneously and outperforms all the other models.

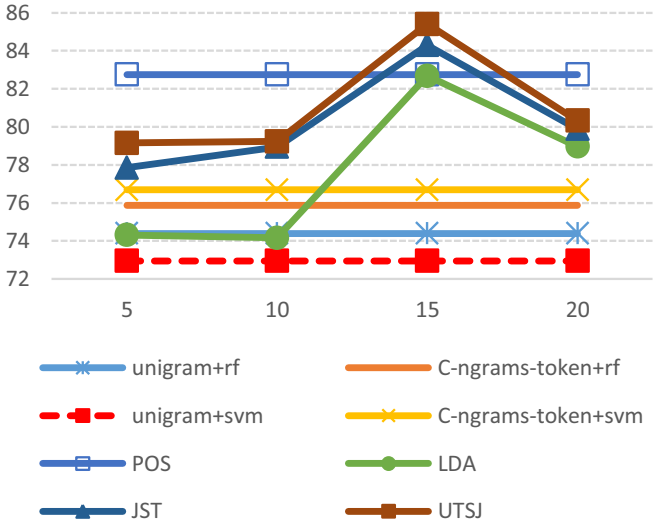
The empirical results of the above experiments supports that our hypothesis "*sentiments dependent upon topic*" is helpful in mining customers' behavioral characteristics of writing reviews. Furthermore, from these results, we can conclude that the sentiment joint topic information can improve the performance of detecting deceptive review. Especially, the performances of all models over unbalanced dataset are correspondingly worse than those over bal-



(a) Precision



(b) Recall



(c) F1

Fig. 8. Performance comparison of the models on unbalanced hotel dataset.

anced dataset, which explains why it is difficult to detect deceptive reviews in real-life electronic ecommerce environment. In contrast to other models, our model have superiority in processing unbalanced dataset, especially good at dealing with big volume of unbalanced dataset (i.e., restaurant unbalanced dataset). This indicates that our model is appropriate to apply in real-life e-commerce environment.

Conclusion 3. In different tested domains, our UTSJ model is superior to benchmark models.

Conclusion 4. Compared to other models, the UTSJ model given in this paper have superiority in dealing with real-life big unbalanced dataset. Therefore, it is more appropriate to apply in real e-commerce environment.

As it can be seen from Table 3 and Table 4, some models (such as POS, LDA, JST, and our UTSJ) may look similar in performance. In order to determine whether the performance differences of our approach and the other methods are statistical significant or not, we calculate the Mann-Whitney *U* test (Mann, & Whitney, 1947) through a two-tailed hypothesis and significance level of 0.05. The null hypothesis states that the models perform equally well with our UTSJ model. Table 5 lists the p-value that are calculated between our UTSJ model and the benchmark models.

From the statistical test results over the hotel dataset, we can see that the POS model and the LDA model gets p-values that is much smaller than 0.05, respectively. This indicates that the differences between our UTSJ model, the POS model and the LDA model are statistically significant, respectively. However, the p-value between JST model and our UTJS model is larger than 0.05, which means that the performance of these two models are not signifi-

Table 5
Statistical significance test of performance.

		Hotel			Restaurant		
		POS	LDA	JST	POS	LDA	JST
UTJS	**Balanced dataset**	1.28e-04	1.21e-06	0.137	6.89e-23	1.30e-14	2.12e-09
	Unbalanced dataset	1.29e-03	0.001	0.232	3.93e-07	3.06e-08	3.28e-25

cantly different over the small hotel dataset, neither balanced nor unbalanced. Differently, all the p-values obtained in the comparison of restaurant corpus is obviously smaller than 0.05, which indicates that our UTSJ model is significantly different to other models. Comprehensive consideration of this result and the high performance of our strategy gotten from Table 4, we can conclude that our UTSJ model is better than POS model, LDA model and JST model in the restaurant corpus. From Table 2, we know that the size of restaurant dataset is much larger than hotel dataset. These results further support above conclusions drawn from samples.

In addition to above experiments, we make some statistics over the outputs of our model. Through qualitative analysis, we find that the truthful reviews tend to express both positive sentiment and negative sentiment towards a specific aspect whereas deceptive reviews tend to express unipolar sentiment towards a specific aspect. Besides, deceptive reviews concentrate on several specific aspects. For example, deceptive restaurant reviews tend to describe food and service whereas truthful restaurant reviews describe various aspects such as location, decoration, parking except for food and service. These findings are byproducts which can further demonstrate that deception is associated with specific sentiments expressed over particular topics. The above statistic results further prove our probabilistic model have superiority in interpretability.

5. Conclusions

This paper proposed a new model named UTSJ, which extends the basic LDA model by adding the sentiment level. We first apply *Gibbs sampling* algorithm to estimate the UTSJ model based on words. Then our new model generates topic-sentiment joint probabilistic distribution through approximating parameters of maximum likelihood function.

Our UTSJ model can mine the topic-sentiment joint information of reviews and obtain the characteristics of reviewers' writing behavior from different topic-sentiment joint probabilistic distributions at the same time. To verify the performance of UTSJ model, we apply it in the deceptive reviews detection over the Yelp multi-domain datasets. According to the results we get from the experiment, we can see that this model not only get better semantics and sentiments of reviews but also can more accurately filter deceptive reviews. Therefore, with the well-trained UTSJ model, electronic commerce platforms can timely capture detection results so that appropriate actions can be further taken (e.g., strengthen the tracking of corresponding reviewers, filtering deceptive reviews, or even downgrading deceptive reviewers' reputation) (Chiu, Leung, & Lam, 2009), and further to information reliable under diasterous conditions (Chiu et al., 2010).

Though this model outperforms traditional methods in statistics and computer linguistics to some extent, this model has still some restrictions that need improve in the future. Firstly, giving this task, our work mainly focuses on obtaining sentiment joint topic probabilistic distribution feature to include this feature vector to the classifier. For example, researchers (Blei & Mcauliffe, 2007) proposed a supervised LDA (sLDA) model that accommodates a variety of response types. So, we can add an extra variable about "deception/truthfulness" to our UTSJ model in order to skip the classifier. Secondly, in the experiments of this paper, the UTSJ model is only compared with feature-based models. Our need phase of work is to verify whether our UTSJ model is better than recent neural network-based representation learning algorithms (Li et al., 2017; Ren & Ji, 2017), as well as regarding other aspects such as good interpretation and high computational efficiency. Thirdly, topic-based summarization of filtered authentic reviews worth further investigation (Condori & Pardo, 2017), as it is helpful for companies to learn customers' feedbacks for improving product design and quality. Moreover, as the resulting topics and corresponding sentiment of authentic reviews can reveal honest reviewers' preferences and opinions about various aspects of the products or services, topic-based recommendation of products and services is a promising research direction (Bauman, Liu, & Tuzhilin, 2017; Bilici & Saygın, 2016). That is because it can depict customers' preference in more detail and generates more appropriate recommendations. Besides, in future work, we will integrate the UTSJ model with the work of Zhao et al. (2015), 2017) and use it in social media text classification.

Acknowledgements

This paper is supported in part by the Natural Science Foundation of China (No. 71772107, 71403151, 61502281, 61433012), the Key R&D Plan of Shandong Province (NO.2018GGX101045), the Natural Science Foundation of Shandong Province (Nos. ZR2018BF013, ZR2013FM023, ZR2014FP011), Shandong Education Quality Improvement Plan for Postgraduate, China's Post-doctoral Science Fund (No. 2014M561948), Postdoctoral innovation project special funds of Shandong Province (No. 201403007), Applied research project for Qingdao postdoctoral researcher, Project of Shandong Province Higher Educational Science and Technology Program (J14LN33), the Leading talent development program of Shandong University of Science and Technology and Special funding for Taishan scholar construction project.

References

Bartcus, M., Chamroukhi, F., & Glotin, H. (2015). Hierarchical Dirichlet Process Hidden Markov Model for unsupervised bioacoustic analysis. In *Paper presented at International Joint Conference on Neural Networks* (pp. 1–7).
Bauman, K., Liu, B., & Tuzhilin, A. (2017). Aspect Based Recommendations: Recommending Items with the Most Valuable Aspects Based on User Reviews. In *Paper presented at The ACM SIGKDD International Conference* (pp. 717–725).
Bilici, E., & Saygın, Y. (2016). Why do people (not) like me?: Mining opinion influencing factors from reviews. *Expert Systems with Applications, 68*, 185–195.
Blei, D. M., Ng, A. Y., & Jordan, M. (2003). Latent dirichlet allocation. *J Machine Learning Research Archive, 3*, 993–1022.
Blei, M. D., & Mcauliffe, J. D. (2007). Supervised topic models. *Advances in Neural Information Processing Systems, 3*, 327–332.
Boyd, R. L., & Pennebaker, J. W. (2016). A Way with Words: Using Language for Psychological Science in the Modern Era. In *Consumer psychology in a social media world* (pp. 222–236). New York, NY, US: Routledge/Taylor & Francis Group.
Cagnina, L., & Rosso, P. (2015). Classification of deceptive opinions using a low dimensionality representation. In *Paper presented at The Workshop on Computational Approaches To Subjectivity* (pp. 58–66).
Chiu, D. K. W., Leung, H. F., & Lam, K. M. (2009). On the making of service recommendations: An action theory based on utility, reputation, and risk attitude. *Expert Systems with Applications, 36*(2), 3293–3301.
Chiu, D. K., Lin, D. T., Kafeza, E., Wang, M., Hu, H., Hu, H., & Zhuang, Y. (2010). Alert based disaster notification and resource allocation. *Information Systems Frontiers, 12*(1), 29–47.
Choi, W. S., & Kim, S. B. (2015). N-gram feature selection for text classification based on symmetrical conditional probability and tf-idf. *Journal of Korean Institute of Industrial Engineers, 41*(4), 381–388.

Condori, R. E. L., & Pardo, T. A. S. (2017). Opinion summarization methods: Comparing and extending extractive and abstractive approaches. *Expert Systems with Applications, 78*, 124–134.
Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.
Deerwester, S. (1988). Improving Information Retrieval with Latent Semantic Indexing. *American Society for Information Science, 100*(1–4), 105–137.
Drummond, C., & Holte, R. C. (2003). C4.5, class imbalance, and cost sensitivity: Why under-sampling beats oversampling. In *Paper presented at Proc of the Icml Workshop on Learning from Imbalanced Datasets II* (pp. 1–8).
Elberrichi, Z. (2006). Text mining using N-grams. *Paper presented at International Conference on Computer Science and ITS Applications*.
Feng, S., Banerjee, R., & Choi, Y. (2012). Syntactic stylometry for deception detection. In *In Paper presented at Meeting of the Association for Computational Linguistics* (pp. 171–175). Short Papers.
Hernández Fusilier, D., Montes-y-Gómez, M., Rosso, P., & Guzmán-Cabrera, R. (2015a). Detection of Opinion Spam with Character n-grams. In *Paper presented at International Conference on Intelligent Text Processing and Computational Linguistics: Vol.9042* (pp. 285–294).
Hernández-Fusilier, D., Montes-Y-Gómez, M., & Rosso, P. (2015b). Detecting positive and negative deceptive opinions using pu-learning. *Information Processing & Management, 51*(4), 433–443.
Ghosh, S., Tonelli, S., & Johansson, R. (2013). Mining fine-grained opinion expressions with shallow parsing. In *In Paper presented at Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013* (pp. 302–310).
Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Paper presented at Fifteenth Conference on Uncertainty in Artificial Intelligence: Vol.41* (pp. 289–296).
Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In *Paper presented at International Conference on Web Search and Data Mining* (pp. 219–230).
Jo, Y., & Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In *Paper presented at ACM International Conference on Web Search and Data Mining: Vol.81* (pp. 815–824).
Johnson, M. J., & Willsky, A. (2012). The Hierarchical Dirichlet Process Hidden Semi–Markov Model. In *Paper presented at Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)* (pp. 252–259).
Johnson, R., & Zhang, T. (2014). Effective use of word order for text categorization with convolutional neural networks. In *Paper presented at Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 103–112).
Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. In *Paper presented at Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 212–217).
Cagnina, L., & Rosso, P. (2017). Detecting deceptive opinions: Intra and cross-domain classification using an efficient representation. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, 25*(Suppl. 2), 151–174.
Li, F., Huang, M., & Zhu, X. (2010). Sentiment analysis with global topics and local dependency. In *Paper presented at Twenty-Fourth AAAI Conference on Artificial Intelligence* (pp. 1371–1376).
Li, J., Cardie, C., & Li, S. (2013). TopicSpam: A Topic-Model based approach for spam detection. In *Paper presented at Meeting of the Association for Computational Linguistics* (pp. 217–221).
Li, J., Dan, J., & Hovy, E. (2015a). When are tree structures necessary for deep learning of representations? In *Paper presented at Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2304–2314).
Li, J., Luong, M. T., & Dan, J. (2015b). A hierarchical neural autoencoder for paragraphs and documents. In *Paper presented at Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing: 1* (pp. 1106–1115).
Li, K., Xie, J., Sun, X., & Bai, H. (2011). Multi-class text categorization based on lda and svm. *Procedia Engineering, 15*(1), 1963–1967.
Li, L., Qin, B., Ren, W., & Liu, T. (2017). Document representation and feature combination for deceptive spam review detection. *Neurocomputing, 254*, 33–41.
Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. *In Paper presented at ACM Conference on Information and Knowledge Management, 217*, 375–384.
Lin, C., He, Y., Everson, R., & Ruger, S. (2012). Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge & Data Engineering, 24*(6), 1134–1145.
Liu, R., Wei, Z., Liu, H., & Fu, Q. Q. (2015). A Part of Speech Based Public Opinion Text Classification Method. *Paper presented at International Conference on Humanities and Social Science Research*.
Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Harvard Business School Working Papers, 62*, 3412–3427.
Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*(1), 50–60.
Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. X. (2007). Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Paper presented at International Conference on World Wide Web* (pp. 171–180).
Mukherjee, S., Dutta, S., & Weikum, G. (2016). Credible Review Detection with Limited Information using Consistency Features. In *Paper presented at Proc. of the Machine Learning and Knowledge Discovery in Databases-European Conference (ECML-PKDD): 9852* (pp. 195–213). LNCS.
Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013). What yelp fake review filter might be doing. *Paper presented at Seventh International AAAI Conference on Weblogs and Social Media*.
Ott, M., Choi, Y., Cardie, C., & Hancock, J. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Paper presented at Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: 1* (pp. 309–319). Human Language Technologies.
Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., & Booth, R.J. (2007). The development and psychometric properties of liwc2007. Austin, 29(11), 1020–1025.
Pennebaker, J. W., Facchin, F., & Margola, D. (2010). What our words say about us: The effects of writing and language. In V. Cigoli, & M. Gennari (Eds.), *Close relationships and community psychology: an international perspective* (pp. 103–117). Milan, Italy: FrancoAngeli.
Ren, Y., & Ji, D. (2017). Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences, 385*, 213–224.
Suess, E. A., & Trumbo, B. E. (2010). Using gibbs samplers to compute bayesian posterior distributions. *Introduction to Probability Simulation and Gibbs Sampling with R*, 219–248.
Sun, C., Du, Q., & Tian, G. (2016). Exploiting product related review features for fake review detection. *Mathematical Problems in Engineering, 1*, 1–7.
Tang, D., Qin, B., & Liu, T. (2015). Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In *Paper presented at Conference on Empirical Methods in Natural Language Processing* (pp. 1422–1432).
Titov, I., & McDonald, R. (2008). Modeling online reviews with multi-grain topic models. In *Paper presented at International Conference on World Wide Web: Vol.41* (pp. 111–120).
Zhang, X. W., & Wu, B. (2016). Short Text Classification based on feature extension using The N-Gram model. In *Paper presented at International Conference on Fuzzy Systems and Knowledge Discovery* (pp. 710–716).
Zhao, Z., Zhang, Y., Li, C., Ning, L., Fan, J., Zhao, Z., et al. (2017). A system to manage and mine microblogging data. *Journal of Intelligent & Fuzzy Systems, 33*(1), 1–11.
Zhao, Z., Li, C., Zhang, Y., Huang, J. Z., Luo, J., Feng, S., et al. (2015). Identifying and analyzing popular phrases multi-dimensionally in social media data. *International Journal of Data Warehousing & Mining, 11*(3), 98–112.