Department

Segment-Level Joint Topic-Sentiment Model for Online Review Analysis

Qinjuan Yang
Sun Yat-sen University

Jiahai Wang
Sun Yat-sen University

Yanghui Rao
Sun Yat-sen University

Fu Lee Wang
Caritas Institute of Higher Education

Haoran Xie
The Education University of Hong Kong

Wai Hong Chan
The Education University of Hong Kong

Editor: Erik Cambria, Nanyang Technological University, Singapore

***Abstract*—With the rapid development of the Internet, an increasing number of users enjoy to shop online and express their reviews on the products and services. Analysis of these online reviews can not only help potential users make rational decisions when purchasing but also improves the quality of products and services. Hence, sentiment analysis for online reviews has become an important and meaningful research domain.**

Unlike traditional sentiment classification of text,[1] which aims only at detecting the sentiment polarities (positive or negative) of documents, the objectives of review analysis also involve extracting the aspects,[2] i.e., the specific features of a product or service that users like or dislike. For example, a movie review that states, "This movie is funny but the ticket is a little expensive," (Review 1) expresses an overall positive sentiment. More specifically, the user expresses a positive sentiment about the "content" and a negative sentiment about the "price." The "content" and "price" are called aspects.

During recent years, many models extended from latent Dirichlet allocation[3] have been effectively applied to review analysis. These models have approached the problem of sentiment analysis at various levels. Some work[4] detects sentiment polarities at the word level, i.e., it assumes

Digital Object Identifier 10.1109/MIS.2019.2899142
Date of current version 12 March 2019.

Table 1. Notations of model variables.

Notations	Description
D	The number of documents
L	The number of sentences in a document
M	The number of segments in a sentence
N	The number of words in a segment
T	The number of topics
V	The vocabulary size
S	The number of sentiment polarities
z	The topic index
s	Sentiment polarity
$\boldsymbol{\theta}$	Multinomial distribution over topic
π	Multinomial distribution over sentiments for each topic
$\boldsymbol{\phi}$	Multinomial distribution over words for each topic and sentiment
$\boldsymbol{\alpha}$	The prior observation count for the frequency of a topic was sampled from a document before any word is observed
$\boldsymbol{\beta}$	The prior observation count for the frequency of word was sampled from a topic and a sentiment before any word is observed
$\boldsymbol{\gamma}$	The prior observation count for the frequency of sentiment polarity was sampled from a document before any word is observed

each word has a sentiment polarity, which is not the case since many words actually do not indicate any sentiment and thus have no effect on the results of sentiment analysis. Other work[5,6] assumes that all words in a sentence indicate the identical sentiment, which is not appropriate to compound and complex sentences that may express more than one sentiment,[7,8] such as the sentence in Review 1.

Still considering Review 1, we can observe that the sentence can be split into two segments by the conjunction "but", i.e., "the movie is funny" and "the ticket is a little expensive." The former segment expresses the positive sentiment whereas the latter shows a negative opinion. Hence, conjunctions play an important role in capturing the sentiment transition among segments.[9,10] Segments connected by coordinating conjunctions (e.g., "and") belong to the same sentiment orientation, whereas segments connected by adversative conjunctions (e.g., "but" and "however") belong to different sentiment orientations.

In this paper, we propose a segment-level joint topic-sentiment model (STSM), in which each sentence is split into segments by conjunctions and the assumption that *all words in a segment express one sentiment* is introduced. It needs to be emphasized that we do not define what sentiment is conveyed when a specific segment appears but estimate the sentiment polarity by our model.

SEGMENT-LEVEL SENTIMENT TOPIC MODEL

The objective of this study is to estimate the sentiment polarity of a document by capturing the topic-sentiment correlation. To model the joint topic-sentiment correlation, a sentiment layer is inserted between the topic layer and the segment layer. Then, we introduce the assumption of "*one segment expresses one sentiment*" to estimate fine-grained sentiments, in which all words in a segment express the same sentiment. For convenience in describing our method, frequently used notations are summarized in Table 1.

Graphical Representation

To model the topic-sentiment correlation and capture the sentiment relative to topics, we introduce a sentiment layer at the segment level. The node of sentiment is inserted after the topic node because the sentiment depends on specific topics. The graphical representation of STSM is shown in Figure 1.

In STSM, words are generated as follows:

1. For each topic z and sentiment s, draw a multinomial word distribution $\phi_{z,s} \sim Dir(\beta)$;
2. For each document d:
   (a) Draw a multinomial topic distribution $\theta_d \sim Dir(\alpha)$;
   (b) For each topic z, draw a multinomial distribution over sentiment polarities $\pi_{d,z} \sim Beta(\gamma)$;
   (c) For each sentence l:
      (i) Draw a topic $z \sim Mul(\theta_d)$;
      (ii) For each segment m, draw a sentiment $s \sim Mul(\pi_{d,z})$;
      (iii) For each word, draw $w \sim Mul(\phi_{z,s})$;

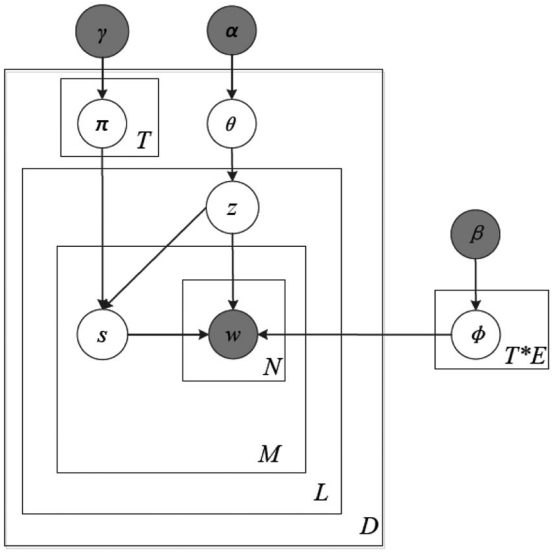

Figure 1. Graphical representation of STSM.

Model Estimation

We use Gibbs sampling to estimate parameters. In each iteration, a topic assignment z for each sentence is drawn from the following conditional probability distribution:

$$p(z_{d,l}=i|z_{-d,l},s,w;\alpha,\beta) \propto \frac{C^{\mathrm{DT}}_{d,i}+\alpha}{\sum_{i'=1}^{T} C^{DT}_{d,i'}+T\alpha} \times \frac{\Gamma\left(\sum_{j'=1}^{E} C^{\mathrm{DTE}}_{d,i,j}+E\gamma\right)}{\Gamma\left(\sum_{j'=1}^{E} C^{\mathrm{DTE}}_{d,i,j}+E\gamma+n_l\right)} \prod_{j=1}^{E} \frac{\Gamma\left(C^{\mathrm{DTE}}_{d,i,j}+\gamma+n_{l,j}\right)}{\Gamma(C^{\mathrm{DTE}}_{d,i,j}+\gamma)} \times \prod_{j=1}^{E} \frac{\Gamma\left(\sum_{k'=1}^{V} C^{\mathrm{TEV}}_{i,j,k'}+V\beta\right)}{\Gamma\left(\sum_{k'=1}^{V} C^{\mathrm{TEV}}_{i,j,k}+V\beta+nw_{l,j}\right)} \times \prod_{k=1}^{V} \frac{\Gamma(C^{\mathrm{TEV}}_{i,j,k}+\beta+nw_{l,j,k})}{\Gamma(C^{\mathrm{TEV}}_{i,j,k}+\beta)},$$

where $z_{d,l}$ is the topic assignment for sentence l in document d; $z_{-d,l}$ is the topic assignment for all sentences expect sentence l in document d; $C^{\mathrm{DT}}_{d,i}$ is the number of sentences assigned to topic i in document d; $C^{\mathrm{DTE}}_{d,i,j}$ is the number of segments assigned to topic i and sentiment j in document d; $C^{\mathrm{TEV}}_{i,j,k}$ is the total number of times word k is assigned to topic i and sentiment j; $n_{l,j}$ is the number of segments assigned to sentiment j in sentence l; n_l is the total number of segments in sentence l; $nw_{l,j,k}$ is the number of times word k is assigned to sentiment j in sentence l; and $nw_{l,j}$ is the total number of times any word is assigned to sentiment j in sentence l.

Table 2. Numbers of sentiment words occurring in each dataset.

Datasets	Lexicon	# of positive words	# of negative words	Total
GameReviews	HSAN	2053	3179	5232
	SAN	3040	2793	5833
PhoneReviews	HSAN	1823	2920	4743
	SAN	2614	2530	5144

The sentiment assignment for each segment is drawn from the conditional probability distribution, as follows:

$$p(s_{d,l,m}=j|s_{-d,l,m},z_{d,l}=i,w;\gamma,\beta) \propto \frac{C^{\mathrm{DTE}}_{d,i,j}+\gamma}{\sum_{j'=1}^{E} C^{\mathrm{DTE}}_{d,i,j'}+E\gamma} \frac{\Gamma\left(\sum_{k'=1}^{V} C^{\mathrm{TEV}}_{i,j,k'}+V\beta\right)}{\Gamma\left(\sum_{k'=1}^{V} C^{\mathrm{TEV}}_{i,j,k}+V\beta+nw_m\right)} \times \prod_{k=1}^{V} \frac{\Gamma(C^{\mathrm{TEV}}_{i,j,k}+\beta+nw_{m,k})}{\Gamma(C^{\mathrm{TEV}}_{i,j,k}+\beta)}$$

where $s_{d,l,m}$ is the sentiment assignment for segment m in sentence l and document d, $s_{-d,l,m}$ is the sentiment assignment for all segments except segment m in sentence l and document d, $nw_{m,k}$ is the frequency of word k occurs in segment m, and nw_m is the total number of words in segment m.

The approximate probability of topic i in document d is

$$\theta_{d,i}=\frac{C^{\mathrm{DT}}_{d,i}+\alpha}{\sum_{i'}^{T} C^{\mathrm{DT}}_{d,i'}+T\alpha}.$$

The approximate probability of sentiment j under topic i in document d is

$$\pi_{d,i,j}=\frac{C^{\mathrm{DTE}}_{d,i,j}+\gamma}{\sum_{j'}^{E} C^{\mathrm{DTE}}_{d,i,j'}+E\gamma}.$$

The approximate probability of word k in topic i with sentiment j is

$$\phi_{i,j,k}=\frac{C^{\mathrm{TEV}}_{i,j,k}+\beta}{\sum_{k'}^{V} C^{\mathrm{TEV}}_{i,j,k'}+V\beta}.$$

For document d, we can estimate its topic distribution $\theta_{d,z}$ and sentiment polarities distribution for a specific topic by $\pi_{d,i,j}$, and the overall

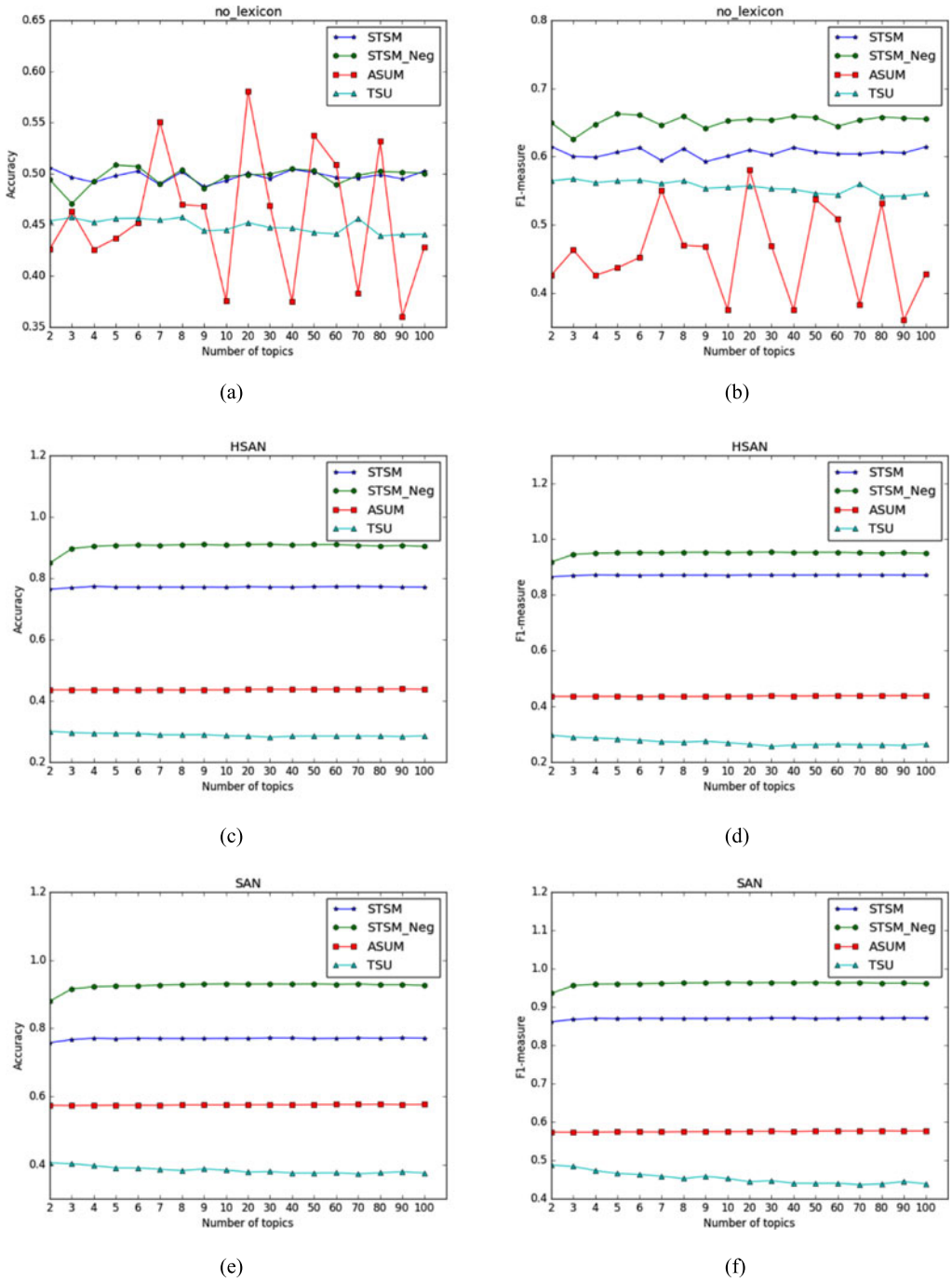

Figure 2. Model performance over GameReviews. (a) Accuracy without sentiment lexicons. (b) F1-measure without sentiment lexicons. (c) Accuracy when using HSAN. (d) F1-measure when using HSAN. (e) Accuracy when using SAN. (f) F1-measure when using SAN.

sentiment polarity of document d is then calculated by

$$p(s = j|d) = \sum_{i=1}^{T} \theta_{d,i} \times \pi_{d,i,j}. \quad (1)$$

EXPERIMENTS

Experimental Setup

We use two review datasets collected from Amazon (www.amazon.com) to evaluate our proposed model. The first dataset, referred to as

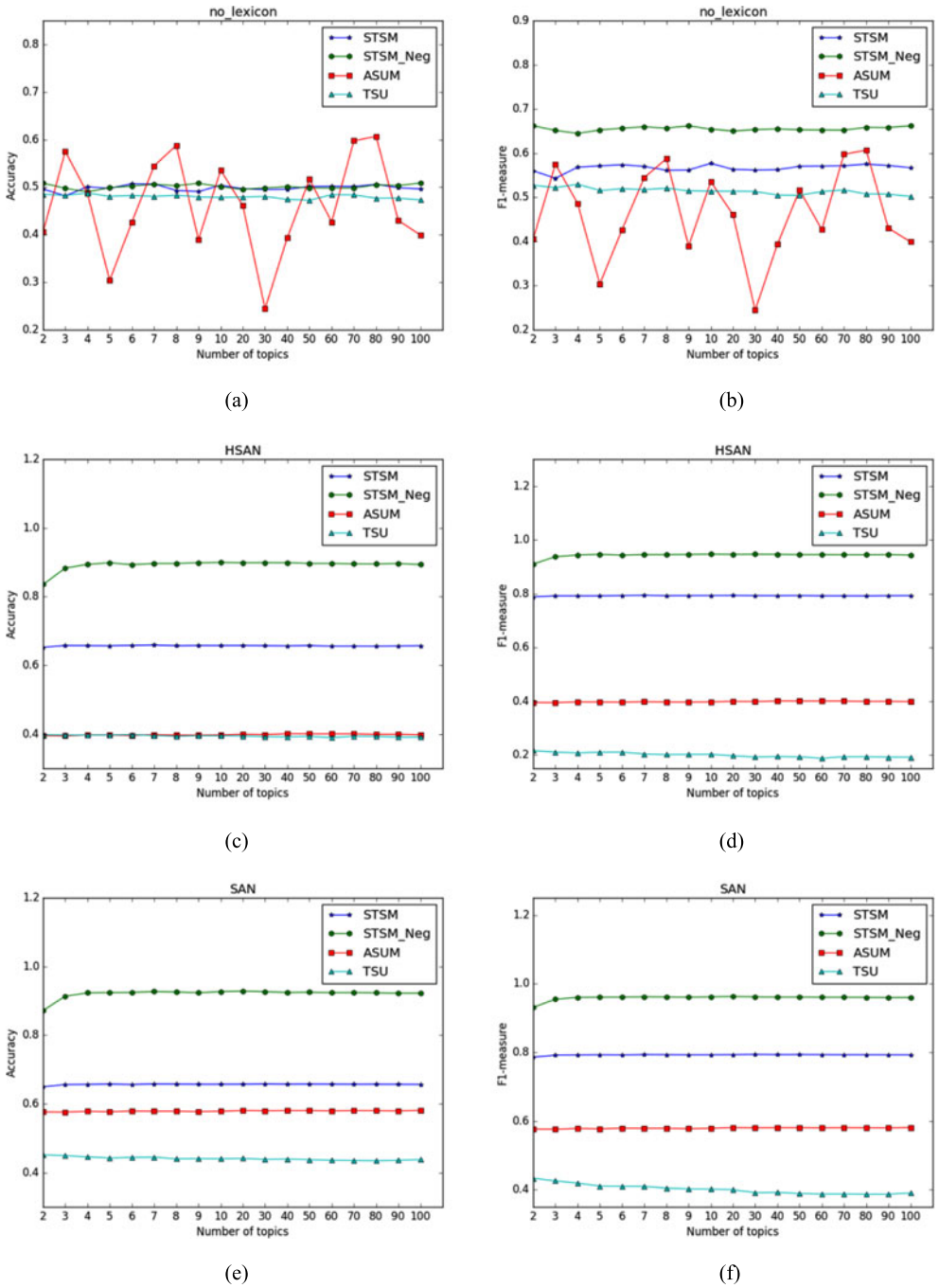

Figure 3. Model performance over PhoneReviews. (a) Accuracy when do not use lexicon.(b) F1-measure when do not use lexicon. (c) Accuracy when use HSAN. (d) F1-measure when use HSAN. (e) Accuracy when use SAN. (f) F1-measure when use SAN.

GameReviews, consists of 2532 reviews of game software and hardware. The other dataset, referred to as *PhoneReviews*, consists of 3265 reviews of phone and phone accessories (headsets and battery chargers). To evaluate the effectiveness of our method, we implemented ASUM[5] and TSU[6] as baselines. Furthermore, to verify the effect of negation on changing the sentiment polarity of text, we also trained the STSM with datasets after negation processing, referred to as STSM_Neg.

We utilized two popular sentiment lexicons named HashtagSentimentAffLexNegLex (HSAN)[11]

Table 3. Several topic and sentiment assignment examples.

Example 1	I have purchased 4 different Jabra headsets because they are the most comfortable and easy to hear with but they have poor microphone quality
STSM_Neg + no lexicon	I have purchased 4 different Jabra headsets because **they are the most comfortable** and ***easy to hear with*** but ***they have poor microphone quality***
STSM_Neg + SAN	I have purchased 4 different Jabra headsets because **they are the most comfortable** and **easy to hear with** but ***they have poor microphone quality***
STSM + no lexicon	I have purchased 4 different Jabra headsets because ***they are the most comfortable*** and **easy to hear with** but ***they have poor microphone quality***
STSM + SAN	I have purchased 4 different Jabra headsets because **they are the most comfortable** and ***easy to hear with*** but ***they have poor microphone quality***
ASUM + no lexicon	I have purchased 4 different Jabra headsets because ***they are the most comfortable*** and ***easy to hear with*** but ***they have poor microphone quality***
ASUM + SAN	I have purchased 4 different Jabra headsets because ***they are the most comfortable*** and ***easy to hear with*** but ***they have poor microphone quality***
TSU + no lexicon	I have purchased 4 different Jabra headsets because ***they are the most comfortable*** and ***easy to hear with*** but ***they have poor microphone quality***
TSU + SAN	I have purchased 4 different Jabra headsets because ***they are the most comfortable*** and ***easy to hear with*** but ***they have poor microphone quality***
Example 2	Not a very good product. Not proportioned to fit my phone even though I had a Frogz brand on my phone already.
	Not a very good product. Not proportioned to fit my phone even though **I had a Frogz brand on my phone already**.
	Not a very good product. Not proportioned to fit my phone even though **I had a Frogz brand on my phone already**.
	Not a very good product. **Not proportioned to fit my phone** even though **I had a Frogz brand on my phone already**.
	Not a very good product. Not proportioned to fit my phone even though **I had a Frogz brand on my phone already**.
	Not a very good product. Not proportioned to fit my phone even though ***I had a Frogz brand on my phone already***.
	Not a very good product. ***Not proportioned to fit my phone*** even though ***I had a Frogz brand on my phone already***.
	Not a very good product. Not proportioned to fit my phone even though ***I had a Frogz brand on my phone already***.
	Not a very good product. ***Not proportioned to fit my phone*** even though ***I had a Frogz brand on my phone already***.

and Sentiment140AffLexNegLex (SAN)[11] to generate positive and negative sentiment words, which are rich enough to cover the vocabularies of the two datasets. Table 2 presents the numbers of sentiment words that occurred in each dataset. For all models, the parameters are set as $\alpha = 50/T$ and $\gamma = 1$. We set the elements of β_s to 0.1 and 0 for those words that occurred and did not occur, respectively, in the word list of sentiment polarity s. The elements of β for words that did not occur in the sentiment lexicon are set to 0.001.

Sentiment Classification

We compare the performance of different models for sentiment classification in terms of accuracy and F1-measure. Figures 2 and 3 show the performance over *GameReviews* and *PhoneReviews*. The results both indicate that STSM_Neg and STSM outperformed ASUM and TSU, which confirms the effectiveness of STSM at detecting segment-level sentiment. Furthermore, we can find that no matter which lexicon was used, the accuracy and F1-measure were much higher and more stable than those of models without lexicons, which validates that

prior knowledge from sentiment lexicons can improve model performance. Moreover, the results of STSM_Neg were always better than STSM, which demonstrates that negation words certainly affect sentiment classification performance.

Topic-Sentiment Alignment

To verify our "*one segment expresses one sentiment*" assumption, several examples for topic and sentiment assignment are presented in Table 3. The results of *PhoneReviews* are shown when no sentiment lexicon and SAN are used. The number of topics is set to 70 to ensure that the model performance has converged. The segments/sentences assigned to positive and negative are marked as bold with underline and bold with italic, respectively. The results show that STSM and STSM_Neg can capture more fine-grained sentiment than ASUM and STU. Furthermore, Example 2 shows that STSM_Neg can effectively handle sentiment detection with negations.

CONCLUSION

In this paper, we present a new approach based on topic models to capture topic-sentiment correlation by drawing topics and sentiments in the topic-sentiment order and assigning topics to sentiment polarities. To capture fine-grained sentiment polarities, we make the assumption "one segment expresses one sentiment" for segment-level sentiment analysis. Experimental results on sentiment classification indicate that the proposed method could boost performance for compound and complex sentences. The effectiveness of our approach is also demonstrated by the alignment of topics and sentiments.

ACKNOWLEDGMENTS

This research was supported in part by the National Natural Science Foundation of China under Grant 61502545, in part by Research Grants Council of Hong Kong Special Administrative Region, China (UGC/FDS11/E03/16), and in part by the Funding Support to ECS Proposal (RG 23/2017-2018R) of The Education University of Hong Kong.

REFERENCES

1. E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.
2. E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," *IEEE Intell. Syst.*, vol. 32, no. 6, pp. 74–80, Nov./Dec. 2017.
3. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, pp. 993–1022, 2003.
4. S. Wang, Z. Y. Chen, and B. Liu, "Mining aspect-specific opinion using a holistic lifelong topic model," in *Proc. 25th Int. Conf. World Wide Web.*, 2016, pp. 167–176.
5. Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, 2011, pp. 815–824.
6. C.L. Ma, M. Wang, and X.W. Chen, "Topic and sentiment unification maximum entropy model for online review analysis," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 649–654.
7. B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.
8. S. Poria, I. Chaturvedi, E. Cambria, and F. Bisio. "Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis," in *Proc. Int. Joint Conf. Neural Netw.*, 2016, pp. 4465–4473.
9. V. Hatzivassiloglou and K.R. McKeown, "Predicting the semantic orientation of adjectives," in *Proc. 8th Conf. Euro. Chapter Assoc. Comput. Linguistics*, 1997, pp. 174–181.
10. S. Poria, E. Cambria, G. Winterstein, and G.B. Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis," *Knowl.-Based Syst.*, vol. 69, pp. 45–63, 2014.
11. S. Kiritchenko, X. Zhu, and S. M. Mohammad, "Sentiment analysis of short informal texts," *J. Artif. Intell. Res.*, vol. 50, pp. 723–762, 2014.

Qinjuan Yang is currently working toward the Master's degree at the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. Contact her at yangqinj@mail2.sysu.edu.cn.

Yanghui Rao is an Associate Professor with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. Contact him at raoyangh@mail.sysu.edu.cn.

Haoran Xie is an Assistant Professor with the Education University of Hong Kong, Hong Kong. Contact him at hxie@eduhk.hk.

Jiahai Wang is a Full Professor with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. Contact him at wangjiah@mail.sysu.edu.cn.

Fu Lee Wang is a Full Professor with the Open University of Hong Kong, Hong Kong. Contact him at pwang@ouhk.edu.hk.

Wai Hong Chan is an Associate Professor with the Education University of Hong Kong, Hong Kong. Contact him at waihchan@eduhk.hk.