TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 09/10 pp528–541
DOI: 10.26599/TST.2019.9010021
Volume 25, Number 4, August 2020

Context-Aware Social Media User Sentiment Analysis

Bo Liu*, Shijiao Tang, Xiangguo Sun, Qiaoyun Chen, Jiuxin Cao, Junzhou Luo, and Shanshan Zhao

Abstract: The user-generated social media messages usually contain considerable multimodal content. Such messages are usually short and lack explicit sentiment words. However, we can understand the sentiment associated with such messages by analyzing the context, which is essential to improve the sentiment analysis performance. Unfortunately, majority of the existing studies consider the impact of contextual information based on a single data model. In this study, we propose a novel model for performing context-aware user sentiment analysis. This model involves the semantic correlation of different modalities and the effects of tweet context information. Based on our experimental results obtained using the Twitter dataset, our approach is observed to outperform the other existing methods in analysing user sentiment.

Key words: social media; sentiment analysis; multimodal data; context-aware; topic model

1 Introduction

Microblogging social networks have become one of the most useful ways for people to express personal opinions and sentiments. Sentiment analysis aims to automatically analyze the user-generated data to discover the sentiments of various users toward products, services, and events[1]. Sentiment analysis is essential for analyzing individual behavior and can be used in several applications, such as forecasting political election results[2], mental health care[3], review analysis[4], and product analysis[5].

Unlike traditional social media, such as newspapers, online social media contains a large amount of multimodal data that can provide a considerably large number of clues for estimating sentiments when compared with that provided by words alone. With the increasing prevalence of smartphones, a growing number of users are inclined to post multimodal messages to express themselves on social network. On the Sina Weibo platform, 95% of the image tweets are accompanied with texts[6], whereas 99% of the image tweets are accompanied with textual content on Twitter[7]. Thus, different modalities in a tweet can be combined for performing sentiment analysis.

Further, the tweets posted on social media usually present abundant contextual information, such as the timelines of the users and the comments of other people. This contextual information is helpful for conducting sentiment analysis because it can comprehensively characterize the contextual attributes of tweet streams. For example, Fig. 1a shows two sequential tweets posted by the same user in an hour. Both the tweets reflect the user's sadness because of the death of Carrier Fisher, indicating that the tweets posted in a short period of time are often sentimentally related. The similar condition can be observed in both the tweets and comments. As depicted in Fig. 1b, the tweet shows a smiling girl with the sentence, "Smile :) it

• Bo Liu, Shijiao Tang, Xiangguo Sun, and Junzhou Luo are with the School of Computer Science and Engineering, Southeast University, Nanjing 211189, China, and also with the Key Laboratory of Computer Network and Information of Ministry of Education of China, Nanjing 211189, China. E-mail: {bliu, sjTang, sunxiangguo, jluo}@seu.edu.cn.

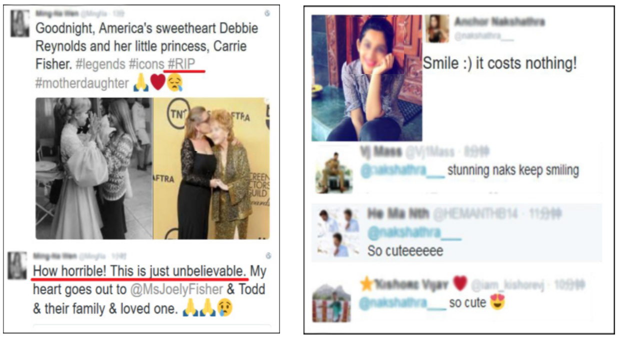
• Qiaoyun Chen is with Microsoft Research Asia, Suzhou 215000, China. E-mail: qychen@seu.edu.cn.

• Jiuxin Cao is with the School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China. E-mail: jx.cao@seu.edu.cn.

• Shanshan Zhao is with the Department of Computer Science and Creative Technologies, University of the West of England, Bristol, BS16 1QY, UK. E-mail: Shanshan.Zhao@uwe.ac.uk.

∗ To whom correspondence should be addressed.

Manuscript received: 2019-04-20; revised: 2019-05-08; accepted: 2019-05-14



[redacted]
Goodnight, America's sweetheart Debbie Reynolds and her little princess, Carrie Fisher. #legends #icons #RIP #motherdaughter 🙏❤️😪

[redacted]
How horrible! This is just unbelievable. My heart goes out to @MsJoelyFisher & Todd & their family & loved one. 🙏🙏😢

(a)

[redacted]
Smile :) it costs nothing!

[redacted]
[redacted] stunning naks keep smiling

[redacted]
[redacted]
So cuteeeeee

★[redacted] ❤️ [redacted]
[redacted] so cute 😍

(b)

Fig. 1 Two types of contextual information in tweets: (a) tweets in users' timelines are related and (b) comments in respond to a tweet are also related.

costs nothing!", reflecting a positive sentiment. This sentiment can be further confirmed by the comments posted by other users with respect to this tweet. A major challenge associated with conducting social media user sentiment analysis is how to model the semantic correlation of different modalities based on the impact of contextual information.

In this study, we use these two types of contextual information along with multimodal data to analyze the semantic correlations that exist among different modalities. Further, we formulate the sentiment factors as latent variables constrained by the sentiment and topic distributions. We subsequently use the probabilistic graphical model technology to characterize the relations that exist between multimodal data and their contextual information. We also propose a sampling algorithm to obtain solutions for our model. The experimental results denote that our model outperforms other methods with respect to sentiment analysis in a multimodal scenario.

The remainder of this paper is as follows. Section 2 provides an overview of the studies conducted in relation to sentiment analysis. Section 3 formulates our sentiment analysis problem and introduces the construction of our Context-Aware Sentiment Analysis (CASA) model. Section 4 presents the sampling algorithms for CASA model inference, and Section 5 illustrates the experimental setup. Extensive experimental results are reported in Section 6. Finally, Section 7 presents the conclusions of our study.

2 Related Work

In this section, we introduce studies related to the visual feature representation of images. Further, we present multimodal sentiment analysis in Section 2.2. Finally, we elaborate on the status of our context-based sentiment analysis research.

2.1 Visual feature representation of images

Images contain several clues for conducting sentiment analysis. One of the most important challenges associated with image sentiment analysis is how to obtain suitable visual features that can reflect the emotions of users. Majority of the existing work conducted in this field is based on low-level visual features[8–11], such as color, texture, and shape. Unfortunately, a considerable affective semantic gap exists between the low-level visual features and the sentiments conveyed by the images. To alleviate this problem, several studies have begun to focus on mid- and high-level visual features. The Bag-of-Visual-Words (BoVW)[12] method maps the key points of images to the visual word vectors that can reveal the characteristics of the images. An alternative model, the principles-of-art-based emotion feature[13] model, unifies various features derived based on different principles such as symmetry, harmony, and gradation. Another model known as Sentribute[14] extracts the low-level features of images and uses a training classifier to generate 102 mid-level attributes to represent these images. In contrast to the aforementioned methods, the Adjective Noun Pair (ANP)[15] method constructs a large-scale visual sentiment ontology to detect the presence of ANPs in an image instead of representing the different features of images.

2.2 Multimodal sentiment analysis

In case of multimodal sentiment analysis, majority of the existing studies have focused on the fusion of multimodal data that includes the following two main methods: early fusion and late fusion. The early fusion method concatenates the textual and visual features into a single feature vector as the input for the sentiment analysis model. The late fusion method first analyzes the textual and visual data, and then combines the output results of different models. Researchers used the early fusion method to generate a joint representation based on different modalities and transmitted it to the downstream classifiers. For example, Wang et al.[16] modeled texts and images in a unified bag-of-words representation and used logistic regression to analyze the sentiments associated with microblogs. Katsurai and Satoh[17] used canonical correlation analysis to project the features of different modalities into a latent embedding space for obtaining a strong correlation among these modalities. You et al.[18] developed a cross-modality regression algorithm to ensure agreement among the sentiment labels predicted using different modality features. Baecchi et al.[19] associated continuous bag-of-words for text feature extraction with a denoising autoencoder for performing image feature extraction and used neural networks to fuse multimodal features for conducting sentiment analysis. Xu and Mao[20] proposed a deep semantic network to combine the features of images and text in a tweet. Although the early fusion methods could capture the correlation among different modalities, the fusion features lacked explicit interpretability, which made the late fusion method a good alternative. The late fusion methods combine the prediction results obtained using several different modalities. For example, Niu et al.[21] proposed a baseline for conducting multimodal sentiment analysis using the late fusion method to combine the analytical results of the textual and visual features. Besides, Cao et al.[22] employed a similar late fusion process to combine the textual and visual sentiment results for conducting sentiment analysis.

The aforementioned methods mainly investigated the fusion approaches for multimodal data, but rarely considered the impact of the contextual information. However, the tweets posted on social media are not isolated and contain abundant contextual information. The contextual information implies the environmental attributes and provides supplemental information for identifying the sentiment associated with a tweet. Further, the short length of the tweets and implicit sentiment words lead to the presence of considerable challenges in understanding the tweet sentiment, which increases the importance of the contextual information for conducting tweet sentiment analysis.

2.3 Context-based sentiment analysis

The contextual information available from tweets characterizes their environmental properties, such as the locations, from which the tweets were tweeted, hashtags, and comments. The contextual information modeling methods comprise matrix factorization and graph models. The matrix-factorization-based methods are superior for mining the latent factors in data. For example, Hu et al.[23] decomposed the message content matrix into user-text and user-user matrices to identify potential relations among different text tweets. They also used the latent relation to conduct sentiment analysis. In another study, Hu et al.[24] used the matrix factorization method to extract emotional clues from the post-word matrix and utilized that information to infer the sentiment labels associated with the posts. Furthermore, Wang et al.[25] proposed a non-negative matrix trifactorization framework to incorporate multiple modalities for identifying the sentiment conveyed by images.

Unfortunately, the computation overhead associated with the matrix factorization methods is often huge. The sparseness of data in social media is also problematic for the application of these methods. In contrast, the probabilistic graphical models can explicitly represent the correlations among different factors and offer an acceptable level of computation time. For example, Yang et al.[8] proposed an emotion learning method by jointly modeling images and comments. Wang et al.[10] considered the impact of social influence and temporal correlation factors on the prediction of the emotional status of users in image-heavy social networks. Yang et al.[11] developed a probabilistic framework to predict the emotional status of various users based on their emotional status histories and social structures in image-based social networks. Vanzo et al.[26] modeled a sequence of tweets related to the same conversation or topic and used SVM$^{\text{HMM}}$ to predict tweet sentiments. Zhao et al.[27] proposed a method to predict the continuous probability distribution of image emotions in the valence arousal space.

Based on the aforementioned discussion, both the multimodal data and contextual information, especially obtained from the users' timelines and comments on their tweets, should be considered for achieving improved sentiment analysis performance. Therefore, we propose a novel CASA model, which will be elaborated in the subsequent section.

3 Sentiment Analysis Model

As mentioned in the previous section, multimodal sentiment analysis rarely considers the contextual information while determining the tweet sentiment. In this section, we propose an unsupervised learning method named CASA to analyze the tweet sentiment using multimodal data based on the contextual information. Our method uses untagged data to mine latent factors from the massive dataset, reducing the cost of manual data tagging. Further, we define the problem in Section 3.1. In Section 3.3, our model is introduced after we presented a set of reasonable hypotheses in Section 3.2.

3.1 Problem statement

Given a user set $U = \{u_1, u_2, \ldots, u_{|U|}\}$, tweet set $D = D_1 \cup D_2 \cup \cdots \cup D_{|U|}$, textual vocabulary $W = \{w_1, w_2, \ldots, w_{|W|}\}$, and visual vocabulary $V = \{v_1, v_2, \ldots, v_{|V|}\}$, we can denote each tweet as $d = \{u, \vec{w}_d, \vec{v}_d, t_d\}$, $d \in D$. Further, $u \in U$ represents the author of d; $\vec{w}_d$ denotes the textual content of d, comprising N_d words selected from the textual vocabulary W; $\vec{v}_d$ is the image content of d, comprising M_d words selected from the visual vocabulary V; and t_d denotes the posting time of d.

Given comments set $R = R_1 \cup R_2 \cup \cdots \cup R_{|U|}$, we use $R_d \subseteq R$ to denote the comment set added to tweet d, and each comment $r \in R_d$ is denoted as a word sequence, the length of this sequence is L_r, and each element (word) can be denoted as a word vector like of $\vec{x}_r$, which is generated from the textual vocabulary W.

Based on the formulation of tweets and comments, the sentiment space and contextual information can be defined as follows:

Definition 1 Sentiment space: A set containing all the possible sentiment values is considered to be the sentiment space. In this study, we define it as $\{$*positive, neutral, negative*$\}$.

Herein, we use the impact of the contextual information associated with a tweet to derive the tweet sentiment; specifically, two types of context information are involved: (1) **users' timelines:** the impact of users' historical sentiment states on the newly posted tweets; and (2) **comments on tweets:** the sentiment correlations that exist among comments and tweets. In formal terms, we can define the tweet contextual information as follows:

Definition 2 Tweet contextual information: For a given user u, we sort all the tweets according to the posting time; for the i-th tweet $d_{u,i}$, we consider the previous tweet $d_{u,i-1}$ and the comment set R_r as contextual information.

Given the aforementioned formulations and definitions, the sentiment analysis problem can be defined as follows. Based on the semantic correlation of different modalities and the impact of contextual information for a given tweet d, we aim to determine the sentiment distribution of d under the sentiment space $\{$*positive, neutral, negative*$\}$.

3.2 Observations and assumptions

The major task of the probability topic model is to discover the correlations among different variables. CASA fuses the latent sentiment factor with the contextual information to construct the generation process of posting tweets and comments. Five hypotheses related to the model are proposed based on the observations of the behavior of posting tweets and comments on social media.

(1) Sentiment labels are associated with topics. The words in different topics may reflect various sentiments[28]. For example, "unpredictable" is negative in "unpredictable steering", but positive in "unpredictable plot". Similarly, depending on the situation, cool-colored images can also express positive ("peaceful") or negative sentiments ("lost and blue"). Thus, we simultaneously model sentiments and topics in this study.

(2) One tweet contains one topic. In social media, one tweet usually contains a single topic. This is caused by the limited character count associated with tweets; thus, tweeting about diverse topics in one short tweet is unrealistic. For instance, tweets have been restricted to 140 characters since the establishment of Twitter, but this character count has been increased to 280 characters from September 2017. Currently, users may use a maximum of four images. Accordingly, majority of the tweets are only related to one obvious topic.

(3) Different modalities exhibit sentiment semantic correlations in the same tweet. In general, the different modalities in a tweet correspond with each other, and tweets are consistent in terms of sentiment expression. Therefore, the text and images in a tweet are assumed to be related to the same topic and exhibit the same sentiment distribution.

(4) Comments can reveal the sentiment of target tweets. The reviewers who post comments under tweets are generally influenced by the sentiment associated with these tweets; the comments are correlated with the corresponding tweets from the sentiment perspective. However, different comments reflect the tweet sentiment to varying extents.

(5) Users' historical tweets in the recent past influence their current sentiment status. Users' sentiments are normally stable over the short term and are highly dependent on their sentiments in the recent past because of the influence of temporal neighborhood information. We assume that the tweets posted in the recent past by a user are sentiment related to construct a correlation among tweets and to ensure model simplicity.

Based on the aforementioned hypotheses, we propose a CASA topic model to describe the generation process of tweets and the corresponding comments. The model exhibits three important characteristics. (1) The sentiment semantic correlation among different modalities is constrained by the overall sentiment distribution and topic in the same tweet. (2) Based on the influence of the temporal neighborhood contextual information, the tweets posted by the same user in the recent past are sentimentally related. (3) The effect of the comment contextual information is considered by introducing a Bernoulli parameter for each comment to bridge the comment to the original tweet.

3.3 Model construction

According to the five hypotheses proposed in the previous subsection, a Bayesian graphical model for sentiment analysis called CASA is conceived by combining the tweets and their contextual information. Figure 2 illustrates the structure of our model, and the notations of the model parameters are presented in Table 1. For the tweet generation process, we use the latent Dirichlet allocation method to develop connections between sentiments and tweets. For obtaining the contextual information, we consider both the comments and the user timelines. The generation process of the proposed model can be elaborated as follows:

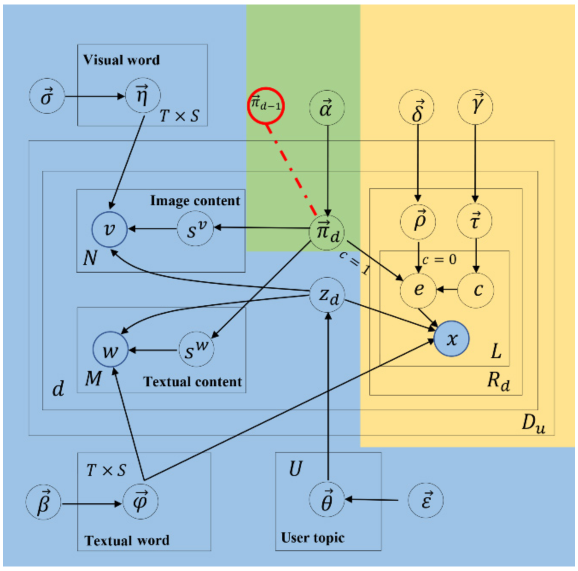


Fig. 2 Graphical representation of the CASA model. The blue and orange blocks describe the generation process of tweets and comments, respectively. The green block describes the correlation among the tweets posted in the recent past.

Table 1 Notation of parameters.

Parameter	Description
Z_d	Topic of tweet
s^w, e	Sentiment label of the textual word
s^v	Sentiment label of the visual word
c	The latent variable that indicates whether the word x in comment r is related to the sentiment of the corresponding tweet d
$\vec{\varphi}_{ks}$	Multinomial distribution in case of textual terms given the topic index k and sentiment index s
$\vec{\eta}_{ks}$	Multinomial distribution in case of visual terms given the topic index k and sentiment index s
$\vec{\theta}_u$	Topic distribution of user u
$\vec{\pi}_d$	Sentiment distribution specific to the tweet d
$\vec{\tau}_r$	Bernoulli distribution over the latent variable c specific to the comment r
$\vec{\rho}_r$	Sentiment distribution specific to the comment r
$\vec{\beta}, \vec{\sigma}, \vec{\varepsilon}, \vec{\alpha}, \vec{\delta}$	Dirichlet priors
$\vec{\gamma}$	Beta prior

Tweet generation process: For each tweet d, the author u initially chooses a topic z_d according to her/his topic distribution $\vec{\theta}$: $z_d \sim \text{Multi}(\vec{\theta})$, where $\vec{\theta}$ is sampled from a Dirichlet distribution with the parameter $\vec{\varepsilon}$. For each textual word w of d, the sentiment label can be generated as $s^w : s^w \sim \text{Multi}(\vec{\pi}_d)$, where $\vec{\pi}_d \sim \text{Dir}(\vec{\alpha})$ is the overall sentiment distribution. For each visual word v of d, the sentiment label can be generated as $s^v : s^v \sim \text{Multi}(\vec{\pi}_d)$. After that, the textual word is generated by $w \sim \text{Multi}(\vec{\varphi}_{z_d, s^w})$, where $\vec{\varphi}_{z_d, s^w}$ is sampled from a Dirichlet distribution with the parameter $\vec{\beta}$. The visual word is generated by $v \sim \text{Multi}(\vec{\varphi}_{z_d, s^v})$, where $\vec{\varphi}_{z_d, s^v}$ is sampled from a Dirichlet distribution with the parameter $\vec{\delta}$. Further, for tweets containing only text or images, we only need to sample the corresponding content. The tweet generation process can be expressed as a joint probability of the topic z_d, tweet sentiment distribution $\vec{\pi}_d$, textual word w_d and its sentiment label $\vec{s^w}_d$, and visual word v_d and its sentiment label $\vec{s^v}_d$:

$$P(\vec{w}_d, \vec{v}_d, \vec{s^w}_d, \vec{s^v}_d, z_d, \vec{\pi}_d, |\vec{\theta}_{u_d}, \Phi, H; \vec{\alpha}) = \\ P(z_d|\vec{\theta}_{u_d}) P(\vec{\pi}_d|\vec{\alpha}) \times \\ \prod_{w \in \vec{w}_d} P(s^w|\vec{\pi}_d) P(w|\vec{\phi}_{z_d, s^w}) \times \\ \prod_{v \in \vec{v}_d} P(s^v|\vec{\pi}_d) P(v|\vec{\eta}_{z_d, s^v}) \quad (1)$$

where $\Phi = \{\vec{\phi}_{11}, \ldots, \vec{\phi}_{|T||S|}\}$ and $H = \{\vec{\eta}_{11}, \ldots, \vec{\eta}_{|T||S|}\}$.

Comment generative process: During the comment generative process, the sentiment label e is sampled from the comment's own sentiment distribution $\vec{\rho}_r \sim \text{Dir}(\vec{\delta})$ or the overall sentiment distribution $\vec{\pi}_d \sim \text{Dir}(\vec{\alpha})$, depending on the situation. We use $\vec{\tau}_r \sim \text{Beta}(\vec{\gamma})$ to represent how likely it is that the sentiment of a comment r is influenced by the sentiment of its corresponding tweet d. Then, a latent variable is sampled as $c \sim \text{Binomial}(\vec{\tau}_r)$, which indicates whether the word is influenced by the sentiment of the corresponding tweet d. If $c = 1$, then the sentiment label e is sampled according to $\vec{\pi}_d$, otherwise, we sample e according to the comment's own sentiment distribution $\vec{\rho}_r$. Finally, the word x is determined as $x \sim \text{Multi}(\vec{\varphi}_{z_d,e})$. The comment generation process can be expressed as the joint probability of the tweet sentiment distribution $\vec{\pi}_d$, topic z_d, $\vec{\tau}_r$, comment sentiment distribution $\vec{\rho}_r$, words in comment $\vec{x}_r$, words' sentiment correlation variable $\vec{c}_r$, and words' sentiment label $\vec{e}_r$.

$$\begin{aligned}
&P(\vec{x}_r, \vec{c}_r, \vec{e}_r, \vec{\rho}_r, \vec{\tau}_r | \vec{\pi}_d, z_d, \Phi; \vec{\delta}, \vec{\gamma}) = \\
&\quad P(\vec{\rho}_r | \vec{\delta}) P(\vec{\tau}_r | \vec{\gamma}) \times \\
&\quad \prod_{x \in \vec{x}_r, c=0} P(c | \vec{\tau}_r) P(e | \vec{\rho}_r) P(x | \vec{\varphi}_{z_d,e}) \times \\
&\quad \prod_{x \in \vec{x}_r, c=1} P(c | \vec{\tau}_r) P(e | \vec{\pi}_d) P(x | \vec{\varphi}_{z_d,e}) \qquad (2)
\end{aligned}$$

Correlation of the adjacent tweets: In Fig. 2 (green block), the correlation between adjacent tweets is denoted using the red dashed line that connects $\vec{\pi}_d$ and $\vec{\pi}_{d-1}$. Based on this, the sequence of $\{\vec{\pi}_1, \vec{\pi}_2, \ldots, \vec{\pi}_{|D_u|}\}$ forms a Markov Random Field (MRF), illustrated in Fig. 3. For any pair of $(\vec{\pi}_i, \vec{\pi}_{i+1})$ in Fig. 3, we define the potential function as

$$\psi(\vec{\pi}_i, \vec{\pi}_{i+1}) = \exp[-l_u \cdot h(t_i, t_{i+1}) \cdot d(\vec{\pi}_i, \vec{\pi}_{i+1})] \quad (3)$$

where

- $d(\vec{\pi}_i, \vec{\pi}_{i+1})$ represents the Euclidean distance between $\vec{\pi}_i$ and $\vec{\pi}_{i+1}$;
- $h(t_i, t_{i+1}) = \mathrm{e}^{(-\omega t_{i+1} - t_i)}$ is the exponential decay function, which describes the time influence, and ω is the decay constant;

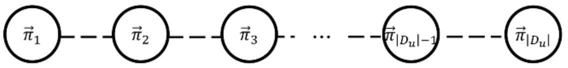


Fig. 3 MRF formed by the sentiment distribution parameters $\vec{\pi}$ of all the tweets D_u posted by a given user u.

- l_u is the user-specific weight parameter, which can describe the user sentiment fluctuation degree.

Further, we place an exponential prior on l_u with the parameter λ,

$$P(l_u | \lambda) = \lambda \mathrm{e}^{-\lambda \cdot l_u} \quad (4)$$

where λ is derived from a Gamma prior with parameters a and b,

$$P(\lambda | a, b) = \frac{b^a \cdot \lambda^{a-1} \cdot \mathrm{e}^{-b\lambda}}{\Gamma(a)} \quad (5)$$

Finally, the joint probability of the model can be deduced as follows:

$$\begin{aligned}
&P(D, R, \vec{z}, \vec{s^w}, \vec{s^v}, \vec{c}, \vec{e}, \Theta, \Phi, H, \Pi, T, P, \vec{l}, \lambda | A) = \\
&\quad \prod_{k \in T} \prod_{s \in S} P(\vec{\varphi}_{ks} | \vec{\beta}) \times \prod_{k \in T} \prod_{s \in S} P(\vec{\eta}_{ks} | \vec{\sigma}) \times \prod_{u \in U} P(\vec{\theta}_u | \vec{\varepsilon}) \times \\
&\quad \prod_{d \in D} P(\vec{w}_d, \vec{v}_v, \vec{S^w}_d, \vec{S^v}_d, z_d, \vec{\pi}_d | \vec{\theta}_{u_d}, \Phi, H; \vec{\alpha}) \times \\
&\quad \prod_{r \in R} P(\vec{x}_r, \vec{c}_r, \vec{e}_r, \vec{\rho}_r, \vec{\tau}_r | \vec{\pi}_d, z_d, \Phi; \vec{\delta}, \vec{\gamma}) \times \\
&\quad \prod_{u \in U} \prod_{d=1}^{|D_u|-1} \Psi(\vec{\pi}_d, \vec{\pi}_{d+1}) \times \prod_{u=1}^{|U|} p(l_u | \lambda) \times p(\lambda | a, b) \\
&\qquad (6)
\end{aligned}$$

where $A = \{\vec{\varepsilon}, \vec{\alpha}, \vec{\beta}, \vec{\sigma}, \vec{\delta}, \vec{\gamma}, \omega, a, b\}$ is the set of hyper-parameters; $\Theta = \{\vec{\theta}_1, \ldots, \vec{\theta}_{|U|}\}$; $\Pi = \{\vec{\pi}_1, \ldots, \vec{\pi}_d, \ldots, \vec{\pi}_{|D|}\}$; $T = \{\vec{\tau}_1, \ldots, \vec{\tau}_{|R|}\}$; $P = \{\vec{\rho}_1, \ldots, \vec{\rho}_r, \ldots, \vec{\rho}_{|R|}\}$.

4 Inference

We infer the sampling formula of the latent variables based on the conjunction relation between the binomial and beta distributions as well as the conjunction relation between the multinomial and Dirichlet distributions. After obtaining the sampling formula, we use the Metropolis-within-Gibbs sampling algorithm[29] to explicitly sample the parameter set $\vec{l}$, $\vec{z}$, $\vec{s^w}$, $\vec{s^v}$, $\vec{c}$, $\vec{e}$, $\vec{\lambda}$, and $\vec{\pi}$. Here, $\vec{\pi}$ is sampled using the Metropolis-Hastings algorithm[30] under the Gibbs sampling framework[31]. The other unknown parameters, including $\vec{\theta}$, $\vec{\varphi}$, $\vec{\eta}$, $\vec{\tau}$, and $\vec{\rho}$, can be obtained from the sampling results.

The sampling rules for the variables are given as follows:

(1) z_d is the topic variable of tweet d.

$$\begin{aligned}
&P(z_d = k | \vec{z}_{\neg d}, D, R, \vec{s^w}, \vec{s^v}, \vec{c}, \vec{e}, \vec{l}, \lambda, \Pi; A) \propto \\
&\quad \frac{n^{(k)}_{u,\neg d} + \varepsilon_k}{\sum_{t \in T} (n^{(t)}_{u,\neg d} + \varepsilon_t)} \times
\end{aligned}$$

$$\prod_{s\in S}\frac{\Gamma(\sum\limits_{w\in W} n_{ks,\neg d}^{(w)}+\beta_w)}{\Gamma(\sum\limits_{w\in W} n_{ks}^{(w)}+\beta_w)}\prod_{w\in W}\frac{\Gamma(n_{ks}^{(w)}+\beta_w)}{\Gamma(n_{ks,\neg d}^{(w)}+\beta_w)}\times$$
$$\prod_{s\in S}\frac{\Gamma(\sum\limits_{v\in V} vn_{ks,\neg d}^{(v)}+\sigma_v)}{\Gamma(\sum\limits_{v\in V} vn_{ks}^{(v)}+\sigma_v)}\prod_{v\in V}\frac{\Gamma(vn_{ks}^{(v)}+\sigma_v)}{\Gamma(vn_{ks,\neg d}^{(v)}+\sigma_v)} \quad (7)$$

where n_u^k is the number of tweets related to topic k posted by user u; $n_{ks}^{(w)}$ is the number of times that the textual term w is assigned to the topic k and sentiment s; $vn_{ks}^{(v)}$ is the number of times that the visual term v is assigned to the topic k and sentiment s; and $\neg d$ denotes a quantity excluding the current instance.

(2) s_i^w is the sentiment variable of the textual word w_i in tweet d.

$$P(s_i^w=l|s^{\vec{w}}_{\neg i},D,R,\vec{z},s^{\vec{v}},\vec{c},\vec{e},\lambda,\vec{l},\Pi;A)\propto$$
$$\pi_{dl}\times\frac{n_{kl,\neg i}^{(w)}+\beta_w}{\sum\limits_{j=1}^{|W|}(n_{kl,\neg i}^{(j)}+\beta_j)} \quad (8)$$

(3) s_j^v is the sentiment variable of the visual word v_i in tweet d.

$$P(s_j^v=l|s^{\vec{v}}_{\neg j},D,R,\vec{z},s^{\vec{w}},\vec{c},\vec{e},\lambda,\vec{l},\Pi;A)\propto$$
$$\pi_{dl}\times\frac{vn_{kl,\neg j}^{(v)}+\sigma_v}{\sum\limits_{i=1}^{|V|}(vn_{kl,\neg j}^{(i)}+\sigma_j)} \quad (9)$$

(4) c_i and e_i are the latent variables of the textual word x_i in comment r specific to tweet d.

$$P(c_i=0,e_i=l|\vec{c}_{\neg i},\vec{e}_{\neg i},D,R,\vec{z},s^{\vec{w}},s^{\vec{v}},\lambda,\vec{l},\Pi;A)\propto$$
$$\frac{n_{r,\neg i}^{(0)}+\gamma_0}{n_{r,\neg i}+\gamma_0+\gamma_1}\times\frac{cn_{r0,\neg i}^{(l)}+\delta_l}{\sum\limits_{s\in S}(cn_{r0,\neg i}^{(s)}+\delta_s)}\times\frac{n_{kl,\neg i}^{(w)}+\beta_w}{\sum\limits_{j=1}^{|W|}(n_{kl,\neg i}^{(j)}+\beta_j)} \quad (10)$$

$$P(c_i=1,e_i=l|\vec{c}_{\neg i},\vec{e}_{\neg i},D,R,\vec{z},s^{\vec{w}},s^{\vec{v}},\lambda,\vec{l},\Pi;A)\propto$$
$$\frac{n_{r,\neg i}^{(1)}+\gamma_1}{n_{r,\neg i}+\gamma_0+\gamma_1}\times\pi_{dl}\times\frac{n_{kl,\neg i}^{(w)}+\beta_w}{\sum\limits_{j=1}^{|W|}(n_{kl,\neg i}^{(j)}+\beta_j)} \quad (11)$$

where $n_r^{(0)}$ and $n_r^{(1)}$ represent the number of times that the latent variable c is sampled to values of 0 and 1, respectively; $cn_{r0}^{(l)}$ is the number of times e assigned to sentiment l in comment r when corresponding c is equal to 0.

(5) l_u is the user-specific weight parameter.

$$P(l_u|\vec{l}_{\neg u},D,R,\vec{z},s^{\vec{w}},s^{\vec{v}},\vec{c},\vec{e},\lambda,\vec{l},\Pi;A)\propto$$
$$\exp\left[-l_u\cdot\left(\lambda+\sum_{d=1}^{|D_u|-1}h(t_d,t_{d+1})\cdot d(\vec{\pi}_d,\vec{\pi}_{d+1})\right)\right] \quad (12)$$

(6) λ is an auxiliary parameter.

$$P(\lambda|D,R,\vec{z},s^{\vec{w}},s^{\vec{v}},\vec{c},\vec{e},\vec{l},\Pi;A)\propto$$
$$\lambda^{a+|U|-1}\cdot\mathrm{e}^{-\lambda(b+\sum\limits_{u\in U}l_u)} \quad (13)$$

(7) $\vec{\pi}_d$ is sentiment distribution parameters for tweet d. We resort to a Metropolis-Hastings step to sample $\vec{\pi}_d$. Given all the current assignments, the proposed distribution can be defined as

$$q(\vec{\pi}_d^*|\vec{\pi}_d^{(t)})\propto\mathrm{Dir}(\vec{\pi}_d^*|\vec{\pi}_d^{(t)}) \quad (14)$$

The acceptance ratio can be derived as

$$\alpha(\vec{\pi}_d^{(t)},\vec{\pi}_d^*)=$$
$$\min\{1,H(\vec{\pi}_d^{(t)},\vec{\pi}_d^*|\Pi_{-d},D,R,\vec{z},s^{\vec{w}},s^{\vec{v}},\vec{c},\vec{e},\vec{l},\lambda;A)\} \quad (15)$$

where $H(\vec{\pi}_d^{(t)},\vec{\pi}_d^*|\Pi_{-d},D,R,\vec{z},s^{\vec{w}},s^{\vec{v}},\vec{c},\vec{e},\vec{l},\lambda;A)$ is the Hastings ratio,

$$H(\vec{\pi}_d^{(t)},\vec{\pi}_d^*|\Pi_{-d},D,R,\vec{z},s^{\vec{w}},s^{\vec{v}},\vec{c},\vec{e},\vec{l},\lambda;A)=$$
$$\exp\left\{l_u\cdot h(t_d,t_{d+1})\cdot[d(\vec{\pi}_d^{(t)},\vec{\pi}_{d+1})-d(\vec{\pi}_d^*,\vec{\pi}_{d+1})]\right\}\times$$
$$\exp\left\{l_u\cdot h(t_d,t_{d-1})\cdot[d(\vec{\pi}_d^{(t)},\vec{\pi}_{d-1})-d(\vec{\pi}_d^*,\vec{\pi}_{d-1})]\right\}\times$$
$$\prod_{s=1}^{S}\left\{(\pi_{ds}^*)^{n_d^{(s)}+\alpha_s-\pi_{ds}^{(t)}}\cdot(\pi_{ds}^{(t)})^{\pi_{ds}^*-n_d^{(s)}-\alpha_s}\cdot\frac{\Gamma(\pi_{ds}^{(t)})}{\Gamma(\pi_{ds}^*)}\right\} \quad (16)$$

Using Eqs. (14) and (15), we can obtain the sampling rule of $\vec{\pi}_d$ at step $t+1$:

(1) Generate a candidate $\vec{\pi}_d^*$ according to Eq. (14);

(2) Calculate the acceptance ratio $\alpha(\vec{\pi}_d^{(t)},\vec{\pi}_d^*)$ using Eq. (15);

(3) Sample a random number $u\sim\mathrm{Uniform}(0,1)$;

(4) If $u<\alpha(\vec{\pi}_d^{(t)},\vec{\pi}_d^*)$, then set $\vec{\pi}_d^{(t+1)}=\vec{\pi}_d^*$, otherwise, set $\vec{\pi}_d^{(t+1)}=\vec{\pi}_d^{(t)}$.

The sampling process includes two stages. In the first stage, burn-in sampling is performed for the first M steps of the total I iterations. In the second stage, $\vec{\pi}_d$ is estimated using the mean value of the results obtained from the remaining $(I-M)$ steps,

$$\vec{\pi}_d=\frac{1}{I-M}\sum_{t=M+1}^{I}\vec{\pi}_d^{(t)} \quad (17)$$

After the sampling process of $\vec{l}, \vec{z}, \vec{s^w}, \vec{s^v}, \vec{c}, \vec{e}, \lambda$, and $\vec{\pi}$, we can use the sampling results as posterior distribution and the prior distribution determined by the hyperparameters to calculate the likelihood parameter Θ, Φ, H, T, and P. The updating rules for Θ, Φ, H, T, and P can be given as follows:

(1) $\vec{\theta}_u$ is the topic distribution specific to user u.

$$\theta_{uk} = \frac{n_u^{(k)} + \varepsilon_k}{\sum\limits_{k' \in T} (n_u^{(k')} + \varepsilon_{k'})} \tag{18}$$

(2) $\vec{\varphi}_{ks}$ denotes the parameter of a multinomial distribution in case of textual terms given the topic index k and sentiment index s.

$$\varphi_{ksw} = \frac{n_{ks}^{(w)} + \beta_w}{\sum\limits_{w' \in W} (n_{ks}^{(w')} + \beta_{w'})} \tag{19}$$

(3) $\vec{\eta}_{ks}$ denotes the parameter of a multinomial distribution in case of visual terms given the topic index k and sentiment index s.

$$\eta_{ksv} = \frac{vn_{ks}^{(v)} + \sigma_v}{\sum\limits_{v' \in V} (vn_{ks}^{(v')} + \sigma_{v'})} \tag{20}$$

(4) $\vec{\tau}_r$ denotes the parameter of the Bernoulli distribution in case of the latent variable c specific to the comment r.

$$\tau_{r0} = \frac{n_r^{(0)} + \gamma_0}{n_r^{(0)} + \gamma_0 + n_r^{(1)} + \gamma_1},$$
$$\tau_{r1} = \frac{n_r^{(1)} + \gamma_1}{n_r^{(0)} + \gamma_0 + n_r^{(1)} + \gamma_1} \tag{21}$$

(5) $\vec{\rho}_r$ denotes the parameter of a sentiment distribution specific to the comment r.

$$\rho_{rs} = \frac{cn_{r0}^{(s)} + \delta_s}{\sum\limits_{s' \in S} (cn_{r0}^{(s')} + \delta_{s'})} \tag{22}$$

The details of our sampling algorithm based on the Metropolis-within-Gibbs sampling method for parameter estimation in CASA are presented in Algorithm 1.

Algorithm 1 Sampling algorithm for CASA model

```
Input: tweet set $D$, comment set $R$, number of topics $|T|$, number of sentiments $|S|$, set of hyper-parameters $A = \{\vec{\varepsilon}, \vec{\alpha}, \vec{\beta}, \vec{\sigma}, \vec{\delta}, \vec{\gamma}, \omega, a, b\}$
Output: latent variables $\vec{z}, \vec{s^w}, \vec{s^v}, \vec{e}, \vec{c}$; parameters of multinomial distributions $\Theta, \Phi, H, \Pi, P$; parameters of binomial distribution $T$; user-specific weight parameters $\vec{l}$; auxiliary parameter $\lambda$
/*initialize */
I = iterations //the predefined times of iterations
$\lambda \sim P(\lambda|a,b)$
for all users $u \in U$ do
    $l_u \sim P(l_u|\lambda)$
for all tweets $d \in D$ do
    $z_d, \vec{s^w}_d, \vec{s^v}_d \sim$ Uniform()
    $\vec{\pi}_d =$ Multi($\frac{1}{s}$)
for all comments $r \in R$ do
    $\vec{c}_r, \vec{e}_r \sim$ Uniform()
/*burn-in and sampling*/
for $i = 1$ to I do
    $\lambda \sim$ Eq. (13)
    for all users $u \in U$ do
        $l_u \sim$ Eq. (12)
    for all tweets $d \in D_u$ do
        $z_d \sim$ Eq. (7)
        generate $\vec{\pi}_d^* \sim$ Eq. (14), calculate $\alpha(\vec{\pi}_d^{(t)}, \vec{\pi}_d^*)$ using Eq. (15), $h \sim$ Uniform(0, 1)
        if $h < \alpha(\vec{\pi}_d^{(t)}, \vec{\pi}_d^*)$ then $\vec{\pi}_d = \vec{\pi}_d^*$
        else $\vec{\pi}_d = \vec{\pi}_d^{(i-1)}$
        for all text words $w_i \in T_d$ and visual words $v_j \in I_d$ do
            $s_i^w \sim$ Eq. (8), $s_j^v \sim$ Eq. (9)
    for all comment $r \in R$ do
        for all text words $x_i \in r$ do
            $c_i, e_i \sim$ Eqs. (10) and (11)
/*update parameters*/
update $\Pi, \Theta, \Phi, H, T, P$ using Eqs. (17)–(22)
```

5 Experimental Setup

5.1 Dataset collection and preprocessing

In this section, we considered Twitter to be our data source for evaluating our model. The dataset used in this study contained all the tweets posted by users in the form of text or image content and all the comments on these tweets. First, we collected the original English tweets posted on May 2014 and the authors' profiles from a Social Media Processing (SMP) 2016 Twitter dataset. Second, we supplemented the images in tweets and their comments using the website crawling method. The users who posted less than 15 tweets were excluded from our final dataset, and the corresponding tweets of these users were also eliminated.

Given the irregularity of tweets in the original dataset, data preprocessing was conducted for both textual and visual contents. On the basis of the unsupervised Bayesian model proposed in this study, these unlabeled data can be directly used for model training after data preprocessing. Table 2 summarizes the basic statistical information obtained in the final dataset after text and image preprocessing. The detailed operations can be

Table 2 Statistics of dataset.

Item	Count
Users	15 765
Tweets	829 296
Image tweets	116 920
Comments	252 668
Textual terms	48 740
Visual terms	750

presented as follows.

5.1.1 Text preprocessing

For text preprocessing, we initially passed the text through an NLTK TweetTokenizer (A famous natural language toolkit, https://www.nltk.org.) to obtain a token list. In social media circumstances, a word containing more than three same letters consecutively exhibits high probability to be an irregular word[32]. For example, some users may use "laaaaaaugh" instead of "laugh" to express their emotions. Therefore, we reduced the repetition length to three if a word containing more than three repeated consecutive letters. Special tokens, such as punctuation marks, URLs, and hashtags, were filtered but emojis and emoticons were retained because of their contributions to sentiment analysis[33]. Subsequently, part-of-speech tagging and named entity recognition were conducted. Spell checking was also conducted using PyEnchant (A spellchecking library for python, https://github.com/rfk/pyenchant.), and stemming work is additionally executed. Subsequently, we applied a frequency filter to omit words that occurred fewer than five times, dropped the stop words, transformed all the words into lower case, and discarded short text-only tweets and comments that contained less than four words. Finally, we obtained 48 740 unique textual words, including 48 099 textual words in lower case and 641 emoji or emoticons.

5.1.2 Image preprocessing

Because topic models were superior for processing discrete data and social media data contained unstructured features, we adopted a Bag-Of-Visual-Words (BOVW) model to transform each image into a bag of emotional words. We initially segmented each image into patches using a graph-based algorithm[34], and subsequently extracted the seven types of features presented in Table 3 for each patch. Additionally, we adopted a z-score method to standardize the features. The k-means method was exploited to construct a visual dictionary. The value of k, also called the size of the dictionary, was determined based on the experimental use of the distortion function. The distortion function was used to calculate the total distance between each instance and the centroid of these instances. The distortion values in case of different numbers of clusters are presented in Fig. 4. Distortion decreased with an increase in the number of clusters. A variation occurred when the number of clusters reached 750; therefore, we set the number of clusters to 750. Finally, each patch was quantified into the closest visual word.

Table 3 Low level visual features for images.

Parameter	Description	Dimension
Hue	Mean and standard deviation of the hue in the HSV color space	2
Saturation	Mean and standard deviation of the saturation in the HSV color space	2
Brightness	Mean and standard deviation of the brightness in the HSV color space	2
Pleasure arousal dominance	Quantitative representation of the image sentiment dimension based on the brightness and saturation in the PAD model[35]	3
Cool color ratio	Colors can be divided into cool colors with a hue value ([0, 360]) between 30 and 110 in the HSV space	1
Local Binary Pattern (LBP)	Uniform pattern of LBP features	10
Coarseness contrast directionality	Three indicators of the Tamura texture features[36]	3

5.2 Model priors definition

To increase the impact of sentiments in CASA model, we add an additional transformation matrix $\boldsymbol{\epsilon}$ to modify the Dirichlet prior $\vec{\beta}$, so that word prior information can be encoded into CASA model according to Ref. [37]. All the elements of $\vec{\beta}$ are initialized with 0.01,

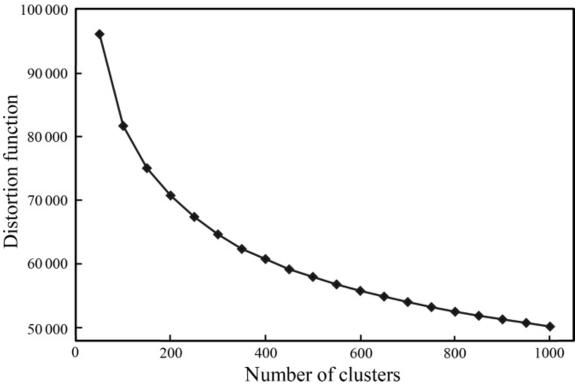


Fig. 4 Distortion function with different numbers of clusters.

and all the elements of $\boldsymbol{\epsilon}$ are initialized with 1. Given a sentiment lexicon SD, each term $w \in W$ and sentiment label $l \in S, S = \{0, 1, 2\}$ (for simplicity, "0" represents "negative", "1" represents "neutral", and "2" represents "positive"), ϵ_{lw} is updated as follows when w occurs in SD,

$$\epsilon_{lw} = \begin{cases} 1, & \text{if } S(w) = l, \\ 0, & \text{otherwise} \end{cases} \tag{23}$$

where $S(w)$ is the prior sentiment label of w in SD. Finally, β_{lw} is updated using $\beta_{lw} = \epsilon_{lw} \times \beta_{lw}$.

On the basis of prior $\vec{\beta}$, the term in sentiment lexicon can only be obtained from the word distribution of the corresponding sentiment. For example, the word "beautiful" with index j in the textual vocabulary occurs in the sentiment lexicon, and its sentiment label is "positive" ($S(w) = l$). The corresponding row in $\boldsymbol{\epsilon}$ is $[0, 0, 1]$, and $\beta_{\cdot j}$ is updated as $\beta_{\cdot j} = [0, 0, 0.01]$. Therefore, "beautiful" can only be obtained from the word distributions specific to a "positive" sentiment. If the term is not present in SD, then $\beta_{\cdot j} = [0.01, 0.01, 0.01]$.

The sentiment prior information was determined based on the textual words and emoji. In case of textual words, the sentiment prior information was extracted from MPQA (MPQA: http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/), and SentiWordNet (SentiWordNet: http://sentiwordnet.isti.cnr.it/). To guarantee the reliability of the sentiment prior, only the words with strong positive or negative orientations in MPQA with a sentiment value larger than 0.7 or smaller than -0.7 in SentiWordNet were extracted. For emojis, the sentiment prior was constructed in Ref. [38], which contained 751 emojis. Considering the emoji lexicon in Ref. [38], we extracted our emoji lexicon according to the following rules:

(1) If the sentiment value is not less than 0.7, we set the prior polarity of this emoji as "positive";

(2) If the sentiment value is not larger than -0.3, we set the prior polarity of this emoji as "negative";

(3) If the ratio of the emoji occurring in negative tweets is less than 0.1, we set the prior polarity of this emoji as "positive" and "neutral".

Based on the aforementioned rules, the prior information statistics are presented in Table 4.

5.3 Parameter configuration

All the hyperparameters of Dirichlet priors, except $\vec{\beta}$ and $\vec{\gamma}$ are symmetric, and the configuration for $\vec{\beta}$ is introduced in Section 5.2. In accordance with the relevant research[8,28,37], other hyper-parameters were set as follows:

- $\epsilon = 50/|T|$,
- $\sigma = 0.01$,
- $\alpha = (0.05 \times L_D)/|S|$ (L_D is the average length of tweets in the corpus),
- $\delta = (0.05 \times L_R)/|S|$ (L_R is the average length of comments in the corpus),
- $\gamma_0 = 1, \gamma_1 = 2$,
- $\omega = 0.8$,
- $a = 2, b = 1$.

Table 4 Statistics of sentiment prior.

Item	Sentiment polarity			
	Positive	Negative	Neutral	Positive/Neutral
Textual words	981	1775	-	-
Emojis	25	7	7	60
Total	1006	1782	7	60

In addition, the configurations for topic number ($|T|$) and iteration times (I) need to be determined through experiments. Here, we used perplexity to set the topic number ($|T|$) and iteration times (I). Perplexity[39] is defined as the reciprocal of the geometric mean of the likelihood of a test corpus. In this study, the perplexity of the CASA model can be defined as follows:

$$\text{Perplexity}(D, R) = \exp\left\{-\frac{\sum\limits_{d \in D} \log P(d)}{\sum\limits_{d \in D} (|T_d| + |I_d| + |R_d|)}\right\} \tag{24}$$

where $P(d)$ is the generating probability of the textual content $\vec{w}_d$, visual content $\vec{v}_d$, and comments R_d of the tweet d,

$$\begin{aligned} P(d) = \sum_{k \in T} \theta_{uk} \sum_{w \in T_d} \sum_{s \in S} \pi_{ds} \varphi_{ksw} + \\ \sum_{v \in I_d} \sum_{s \in S} \pi_{ds} \eta_{ksv} + \\ \sum_{r \in R_d} \sum_{x \in r} \sum_{s \in S} \pi_{ds} \eta_{ksv} \end{aligned} \tag{25}$$

To set the value of I and $|T|$, we initially set $I = 1000$ experimentally and calculated the perplexity of the CASA model with different $|T|$. The perplexity values with different numbers of topics are presented in Fig. 5. As the number of topics increased, the perplexity of CASA decreased until it became flat at 25 topics. Therefore, we fixed $|T| = 25$ and calculated the perplexity of the CASA model at different iteration times. The perplexity values in case of different iteration times are depicted in Fig. 6. When I became

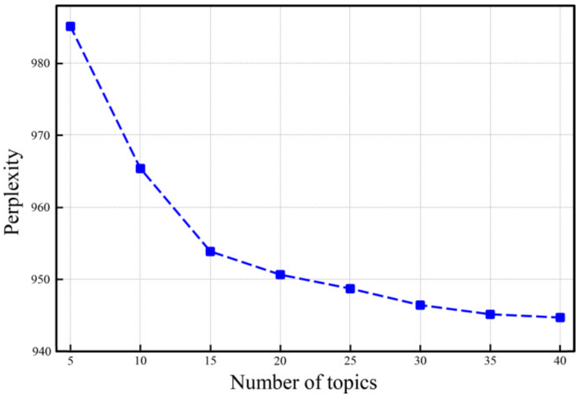


Fig. 5 Perplexity values in case of different numbers of topics.

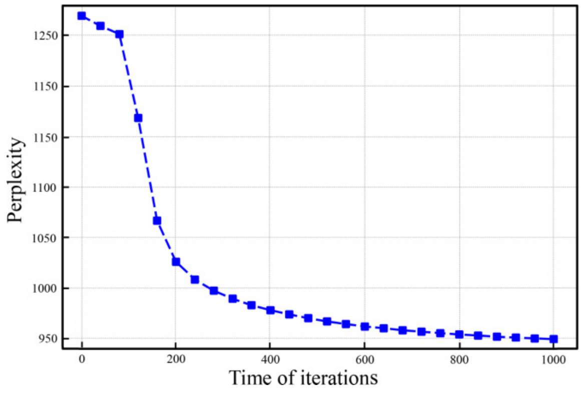


Fig. 6 Perplexity values in case of different times of iterations (the number of topics is fixed as 25).

400, the perplexity became flat; thus, we set $I = 1000$ and saved the sampling results every 40 steps after 400 iterations.

6 Experimental Results and Analysis

6.1 Sentiment annotations

To evaluate the sentiment classification performance, we manually labeled a portion of the tweets' sentiment. First, 1000 tweets containing both text and images were randomly selected from the experimental dataset. Second, these tweets were manually labeled using the sentiment set {*positive, neutral, negative*}. The final sentiment label, namely ground truth, was determined by the majority sentiment polarity results from related tweets. To ensure the reliability of the results, we only retained tweets with a voting proportion of more than 80%. Finally, we obtained the final labeled test dataset containing 456 tweets, including 225 positive tweets, 111 neutral tweets, and 120 negative tweets.

6.2 Comparison algorithms

In this subsection, we select five comparison methods in the field of social media sentiment analysis. The details of these methods are presented as follows.

CASA-reply: The CASA model proposed in this study was used without considering the influence of comments. We used the model to prove the efficiency of the comment context.

CASA-time: The CASA model proposed in this study was used without considering the influence of users' timelines. We used the model to prove the efficiency of this kind of contextual information.

SentiStrength[40]: We used a text sentiment analysis algorithm based on the sentiment lexicon, which is extensively used for short text sentiment detection in social media. We used this method to compare the efficiency of jointly modeling text and images.

SentiBank[15]: As an attribute representation designed for human affective computing, SentiBank includes 1200 ANPs, such as "cloudy moon" and "beautiful rose", which are carefully selected from the web data and represent human effects. SentiBank is intuitively suitable for conducting visual sentiment analysis. We used this method to compare the efficiency of jointly modeling text and images with the CASA model.

T-V-Early[21]: As a baseline for multimodal (text and image) sentiment analysis, this model uses GIST, LBP, and other feature extraction methods to represent the visual features and TF-IDF to represent the textual features. After feature extraction, three methods, including early fusion, late fusion, and Deep Boltzmann Machine (DBM), were used to detect the tweet sentiment. Previous experimental results show that early fusion is better than late fusion and the DBM with respect to the Twitter dataset. Therefore, we used the early fusion strategy and considered the contextual information, so that we can see the performance of our method compared with other models.

6.3 Results and analysis

The output of our CASA model is a 3-dimensional vector $\vec{\pi}_d$, where each entry indicates one of the sentiment polarities of tweet d. The final polarity depends on the entry with the maximal value,

$$\text{Polarity}(d) = \arg\max_{s \in \{neg,\ neu,\ pos\}} \pi_{ds} \tag{26}$$

In this study, the evaluation metrics included accuracy, macro-recall, macro-precision, and macro-F_1[41]. These metrics are commonly used to measure the performance of multiclassification problems.

6.3.1 Model performance

To evaluate the efficiency of the CASA model, we used the preprocessed test dataset described in Section 5 and compared the CASA algorithm with the remaining competing algorithms introduced in Section 6.2 based on the four previously illustrated evaluation metrics.

Figure 7 denotes the performance of the related algorithms based on which it can be stated that the CASA model clearly outperformed the others. With respect to accuracy, our model surpassed the second best technique by approximately 2.8%. In case of macro-precision, CASA performed 4.3% better than the second ranked one. Similar results were observed in case of macro-recall and macro-F_1 with 8.3% and 7.1% improvement, respectively, when compared with the second best technique. SentiBank performed the worst in general because of its limited image feature extraction capability, especially when the conformity between the images and texts was not explicit. T-V-Early performed better than SentiStrength, indicating that **multiple modalities are beneficial for conducting sentiment analysis.** The results reported from T-V-Early and CASA demonstrated that the contextual information of tweets could help improve the sentiment detection performance.

6.3.2 Context contribution analysis

As previously mentioned, two types of contextual information were considered in our model. In this section, we analyzed the contribution of these two types of contextual information when analyzing the tweet sentiment. As depicted in Fig. 8, CASA-reply and CASA-time performed worse than CASA. CASA-time outperformed CASA-reply, which indicated that the comment contextual information is more important than users' timelines in tweet sentiment analysis. The reason for the above result may be twofold. On the one hand, this observation can be attributed to the short length of tweets and the lack of explicit sentiment words in these tweets. However, by integrating the comments and tweets, we alleviated the shortcomings of the limited length and the lack of explicit sentiment words. On the other hand, the CASA model considered the correlations of tweets sent in the recent past, but the sentiment correlation of these adjacent tweets was sometimes slightly weak, which led to the result of CASA-reply being slightly worse than CASA-time.

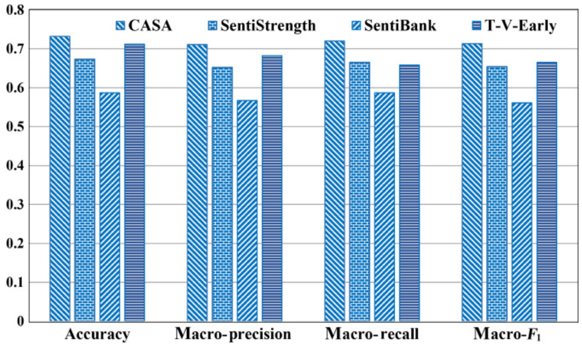


Fig. 7 Performance comparison of different algorithms.

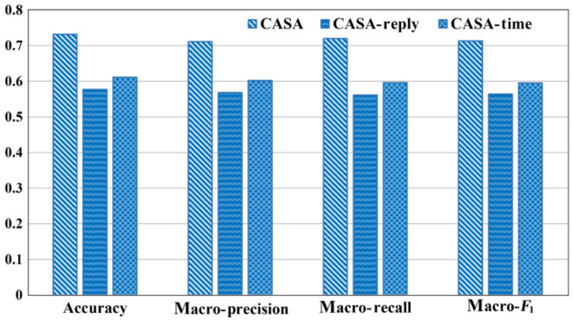


Fig. 8 Performance comparison based on the contributions of comments and users' timelines.

7 Conclusion

In this study, we investigated the problem of social media sentiment analysis. A probability model called CASA is proposed for tweet sentiment analysis in which the semantic correlation of different modalities and the influence of the tweet contextual information are both considered. Through the comparison and analysis of the experimental results, the proposed CASA model was observed to efficiently detect the sentiments contained in multimodal tweets; both the types of contextual information used in our model can significantly improve the model performance.

Acknowledgment

This work was supported by the National Key R&D Program of China (No. 2017YFB1003000), the National Natural Science Foundation of China (Nos. 61972087 and 61772133), the National Social Science Foundation of China (No. 19@ZH014), Jiangsu Provincial Key Project (No. BE2018706), the Natural Science Foundation of Jiangsu Province (No. SBK2019022870), Jiangsu Provincial Key Laboratory of Network and Information Security (No. BM2003201), and Key Laboratory of Computer Network and Information Integration of Ministry of Education of China (No. 93K-9).

References

[1] B. Liu, Sentiment analysis and opinion mining, *Synthesis Lectures on Human Language Technologies,* vol. 5, no. 1,

pp. 1–167, 2012.

[2] D. Paul, F. Li, M. K. Teja, X. Yu, and R. Frost, Compass: Spatio temporal sentiment analysis of US Election what Twitter says!, in *Proc. of the 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Halifax, Canada, 2017, pp. 1585–1594.

[3] M. De Choudhury, S. Counts, E. J. Horvitz, and A. Hoff, Characterizing and predicting postpartum depression from shared Facebook data, in *Proc. of the 17th ACM Conf. on Computer Supported Cooperative Work & Social Computing*, Baltimore, MD, USA, 2014, pp. 626–638.

[4] Z. Zhai, H. Xu, and P. Jia, An empirical study of unsupervised sentiment classification of Chinese reviews, *Tsinghua Science and Technology*, vol. 15, no. 6, pp. 702–708, 2010.

[5] Q. Ye, Y. Li, and Y. Zhang, Semantic-oriented sentiment classification for Chinese product reviews: An experimental study of book and cell phone reviews, *Tsinghua Science and Technology*, vol. 10, no. S1, pp. 797–802, 2005.

[6] C. Tao, Analyzing image tweets in Microblogs, Ph.D. dissertation, School of Computer, National University Of Singapore, Singapore, 2016.

[7] T. Chen, D. Lu, M-Y. Kan, and P. Cui, Understanding and classifying image tweets, in *Proc. of the 21st ACM Int. Conf. on Multimedia*, Barcelona, Spain, 2013, pp. 781–784.

[8] Y. Yang, J. Jia, S. Zhang, B. Wu, Q. Chen, J. Li, C. Xing, and J. Tang, How do your friends on social media disclose your emotions?, in *Proc. of the Twenty-Eighth AAAI Conf. on Artificial Intelligence*, Austin, TX, USA, 2014, pp. 1–7.

[9] Y. Yang, P. Cui, W. Zhu, H. V. Zhao, Y. Shi, and S. Yang, Emotionally representative image discovery for social events, in *Proc. of Int. Conf. on Multimedia Retrieval*, Glasgow, UK, 2014, p. 177.

[10] X. Wang, J. Jia, J. Tang, B. Wu, L. Cai, and L. Xie, Modeling emotion influence in image social networks, *IEEE Transactions on Affective Computing,* vol. 6, no. 3, pp. 286–297, 2015.

[11] Y. Yang, J. Jia, B. Wu, and J. Tang, Social role-aware emotion contagion in image social networks, in *Proc. of the Thirtieth AAAI Conf. on Artificial Intelligence*, Phoenix, AZ, USA, 2016, pp. 65–71.

[12] J. Yang, Y-G. Jiang, A. G. Hauptmann, and C-W. Ngo, Evaluating bag-of-visual-words representations in scene classification, in *Proc. of the Int. Workshop on Multimedia Information Retrieval*, Augsburg, Germany, 2007, pp. 197–206.

[13] S. Zhao, Y. Gao, X. Jiang, H. Yao, T-S. Chua, and X. Sun, Exploring principles-of-art features for image emotion recognition, in *Proc. of the 22nd ACM Int. Conf. on Multimedia*, Orlando, FL, USA, 2014, pp. 47–56.

[14] J. Yuan, S. Mcdonough, Q. You, and J. Luo, Sentribute: Image sentiment analysis from a mid-level perspective, in *Proc. of the Second Int. Workshop on Issues of Sentiment Discovery and Opinion Mining*, Chicago, IL, USA, 2013, p. 10.

[15] D. Borth, R. Ji, T. Chen, T. Breuel, and S-F. Chang, Large-scale visual sentiment ontology and detectors using adjective noun pairs, in *Proc. of the 21st ACM Int. Conf. on Multimedia*, Barcelona, Spain, 2013, pp. 223–232.

[16] M. Wang, D. Cao, L. Li, S. Li, and R. Ji, Microblog sentiment analysis based on cross-media bag-of-words model, in *Proc. of Int. Conf. on Internet Multimedia Computing and Service*, Xiamen, China, 2014, p. 76.

[17] M. Katsurai and S. Satoh, Image sentiment analysis using latent correlations among visual, textual, and sentiment views, in *2016 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 2837–2841.

[18] Q. You, J. Luo, H. Jin, and J. Yang, Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia, in *Proc. of the Ninth ACM Int. Conf. on Web Search and Data Mining*, San Francisco, CA, USA, 2016, pp. 13–22.

[19] C. Baecchi, T. Uricchio, M. Bertini, and A. Del Bimbo, A multimodal feature learning approach for sentiment analysis of social network multimedia, *Multimedia Tools and Applications,* vol. 75, no. 5, pp. 2507–2525, 2016.

[20] N. Xu and W. Mao, MultiSentiNet: A deep semantic network for multimodal sentiment analysis, in *Proc. of the 2017 ACM on Conf. on Information and Knowledge Management*, Singapore, 2017, pp. 2399–2402.

[21] T. Niu, S. Zhu, L. Pang, and A. El Saddik, Sentiment analysis on multi-view social data, in *Int. Conf. on Multimedia Modeling*, Miami, FL, USA, 2016, pp. 15–27.

[22] D. Cao, R. Ji, D. Lin, and S. Li, A cross-media public sentiment analysis system for microblog, *Multimedia Systems,* vol. 22, no. 4, pp. 479–486, 2016.

[23] X. Hu, L. Tang, J. Tang, and H. Liu, Exploiting social relations for sentiment analysis in microblogging, in *Proc. of the Sixth ACM Int. Conf. on Web Search and Data Mining*, 2013, Rome, Italy, pp. 537–546.

[24] X. Hu, L. Tang, J. Tang, H. Gao, and H. Liu, Unsupervised sentiment analysis with emotional signals, in *Proc. of the 22nd Int. Conf. on World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 607–618.

[25] Y. Wang, Y. Hu, S. Kambhampati, and B. Li, Inferring sentiment from web images with joint inference on visual and social cues: A regulated matrix factorization approach, in *Ninth Int. AAAI Conf. on Web and Social Media*, Oxford, UK, 2015, pp. 473–482.

[26] A. Vanzo, D. Croce, and R. Basili, Inferring sentiment from web images with joint inference on visual and social cues: A regulated matrix factorization approach, in *Proc. of COLING 2014, the 25th Int. Conf. on Computational Linguistics: Technical Papers*, Dublin, Ireland, 2014, pp. 2345–2354.

[27] S. Zhao, H. Yao, Y. Gao, R. Ji, W. Xie, X. Jiang, and T-S. Chua, Predicting personalized emotion perceptions of social images, in *Proc. of the 2016 ACM on Multimedia Conf.*, Amsterdam, The Netherlands, 2016, pp. 1385–1394.

[28] C. Lin and Y. He, Joint sentiment/topic model for sentiment analysis, in *Proc. of the 18th ACM Conf. on Information and Knowledge Management*, Hong Kong, China, 2009, pp. 375–384.

[29] G. O. Roberts and J. S. Rosenthal, Examples of adaptive MCMC, *Journal of Computational and Graphical Statistics,* vol. 18, no. 2, pp. 349–367, 2009.

[30] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan,

An introduction to MCMC for machine learning, *Machine Learning,* vol. 50, nos. 1&2, pp. 5–43, 2003.

[31] P. Resnik and E. Hardisty, Gibbs sampling for the uninitiateds, Report, Institute for Advanced Computer Studies, Unveristy of Maryland, College Prk, MD, USA, 2010.

[32] B. Han and T. Baldwin, Lexical normalisation of short text messages: Makn sens a# twitter, in *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, Portland, OR, USA, 2011, pp. 368–378.

[33] T. Hu, H. Guo, H. Sun, T. T. Nguyen, and J. Luo, Spice up your chat: The intentions and sentiment effects of using emoji, in *Proc. of the Eleventh Int. AAAI Conf. on Web and Social Media*, Montreal, Canada, 2017, pp. 102–111.

[34] P. F. Felzenszwalb and D. P. Huttenlocher, Efficient graph- based image segmentation, *International Journal of Computer Vision,* vol. 59, no. 2, pp. 167–181, 2004.

[35] P. Valdez and A. Mehrabian, Effects of color on emotions, *Journal of Experimental Psychology: General*, vol. 123, no. 4, p. 394, 1994.

[36] H. Tamura, S. Mori, and T. Yamawaki, Textural features corresponding to visual perception, *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 8, no. 6, pp. 460–473, 1978.

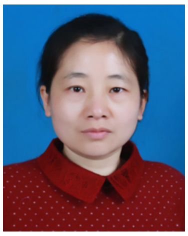
[37] C. Lin, Y. He, R. Everson, and S. Ruger, Weakly supervised joint sentiment-topic detection from text, *IEEE Transactions on Knowledge and Data Engineering,* vol. 24, no. 6, pp. 1134–1145, 2012.

[38] P. K. Novak, J. Smailović, B. Sluban, and I. Mozetič, Sentiment of emojis, *PloS One,* vol. 10, no. 12, p. e0144296, 2015.

[39] H. Gregor, Parameter estimation for text analysis, Report, Fraunhofer Institute for Computer Graphics Research, Darmstadt, Germany, 2005.

[40] M. Thelwall, K. Buckley, and G. Paltoglou, Sentiment strength detection for the social web, *Journal of the American Society for Information Science and Technology,* vol. 63, no. 1, pp. 163–173, 2012.

[41] Z. Zhou, *Machine Learning*. Beijing, China: Tsinghua University Press, 2016.



Bo Liu received the PhD degree from Southeast University in 2007. She is currently an associate professor at the School of Computer Science in Southeast University, Nanjing, China. Her research interests include spammer detection in social network, the evolution of social community, social influence, and multi-agent technology.

Shijiao Tang is currently a master student at the School of Computer Science and Technology in Southeat University, Nanjing, China. His research interests include event detection, event evolution, and sentiment analysis in social media.

Xiangguo Sun is currently a PhD candidate at the School of Computer Science and Technology, Southeast University, Nanjing, China. His research interests include social media analysis, user behaviors mining, network embedding, and sentiment analysis.

Qiaoyun Chen received the master degree from the School of Computer Science and Technology, Southeast University, Nanjing, China. She now works as a research assistant in Microsoft Research Asia. Her research interests include social media analysis, big data analysis, social influence, and user behavior modeling.

Jiuxin Cao received the PhD degree from Xi'an Jiaotong University in 2003. He is currently a professor at the School of Cyber Science and Engineering in Southeast University, Nanjing, China. His research interests include computer networks, social computing, behavior analysis, and big-data security and privacy preservation.

Junzhou Luo received the PhD degree from Southeast University in 2000. He is currently a full professor at the School of Computer Science and Engineering, Southeast University, Nanjing, China. His research interests include next-generation networks, protocol engineering, network security, cloud computing, and wireless LAN.

Shanshan Zhao received the PhD degree from Xi'an Jiaotong University in 2008. She is currently a research fellow at the Faculty of Engineer and Technology, University of West England, Bristol, UK. Her research interests include industrial IoT, optimization algorithms, multiscale modelling, and lightweight computational algorithm.