
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 115–126, DOI 10.12694/scpe.v25i1.2279

AN EMOTIONAL ANALYSIS OF KOREAN TOPICS BASED ON SOCIAL MEDIA BIG DATA CLUSTERING

YANHONG JIN*

Abstract. An innovative approach is introduced in this paper to address the challenges in emotional topic interpretation and accuracy in emotional situation assessment. Utilizing large data from social media to improve the accuracy of emotional analysis in online debates, with a specific emphasis on Korean themes. The proposed solution, the Online Topic Emotion Recognition Model (OTSRM), builds upon the foundational Online Latent Dirichlet Allocation (OLDA) model. The OTSRM integrates the concept of emotion intensity and introduces an inventive emotion iteration framework to tackle these issues. Key innovations of the OTSRM include establishing an affective evolution channel by augmenting affective heritability using a β priori. Additionally, the model generates two critical distribution matrices: one for characteristic words and another for affective words, facilitating a deeper understanding of emotional context within topics. The relative entropy method is employed to discern emotional tones in textual content, calculating maximum emotion values for topic focus within adjacent time segments. Validation experiments using five diverse network event datasets and comparisons to mainstream models demonstrate the OTSRM's effectiveness with emotion recognition accuracy rates of 85.56% and 81.03%. The OTSRM represents significant progress in addressing challenges associated with emotional topic analysis and precise emotional dynamics assessment in Korean social media data.

Key words: Emotional Analysis, Social Media, Topic Emotion Recognition, Affective Evolution, Online Latent Dirichlet Allocation, Public Opinion Analysis

AMS subject classifications. 15A15, 15A09, 15A23

1. Introduction.

The Internet has gained widespread popularity recently, and social media has progressively assumed a significant role in people's daily lives. The real-time recording of users' online activities in the digital world has accumulated a wealth of user behaviour data within natural contexts. This growing data source has accompanied a fresh paradigm and investigative avenue for research. Social media platforms, owing to their diversity of information and complex functionalities, offer researchers many data types to explore. During the investigation, individuals often delve deeply into specific data categories for analysis and practical application. These social media data can be categorized based on their recording forms, encompassing personal account details, usage patterns, textual content, social network interactions, visual content, and other relevant cues [16].

The volume of social media data centered on Korean topics is unparalleled, boasting various informational formats. Consequently, connecting the potential of Korean topic social media big data necessitates comprehensive data collection platforms and a broad-ranging approach to data acquisition. Building upon existing social media intelligence sources, expanding the intelligence sources as extensively as possible is imperative. Concurrently, there's a pressing need to fortify the capabilities of existing data mining technologies, ensuring that recognition techniques transcend the confines of plain text and extend into the domains of images, audio, and video. This evolution aims to enable timely and exhaustive intelligence gathering about Korean topics within social media [9].

In the era of big data, the digital world is overflowing with a creative expression of interactive content driven by public sentiment, giving rise to a multifaceted tapestry of opinion evolution patterns. This torrent of digital discourse encompasses a kaleidoscope of perspectives, reactions, and dialogues, mirroring the dynamic and ever-shifting nature of public sentiment within the online realm. Particularly noteworthy are categories of public sentiment data, such as breaking news and trending events, which wield considerable influence over

*School of Foreign Studies, Lingnan Normal University, Zhanjiang, Guangdong, 524048 China (Corresponding author: yanhongjin3@163.com)

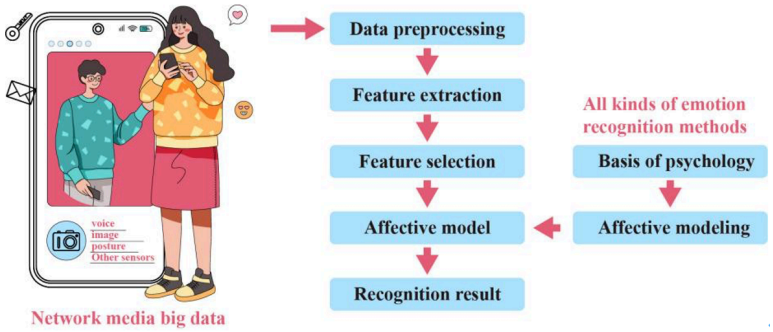

Fig. 1.1: Social media big data clustering

the trajectory of public opinion. These influential occurrences can frequently provoke polarized sentiments, inciting conspicuous shifts in the digital landscape toward either a positive or negative direction. These shifts sometimes escalate into full-fledged online public opinion crises characterized by fervent debates, misinformation proliferation, and emotionally charged exchanges [20].

The far-reaching significance of efforts transcends the confines of data analysis. They serve as vital for establishing effective early warning systems attuned to the fluctuations of public sentiment. By remaining vigilant to the evolving emotional topography of online discourse, it becomes feasible to discern the telltale signs of impending public sentiment crises and proactively implement measures to mitigate their impacts. Ultimately, the overarching goal is to nurture a harmonious and constructive online public opinion environment that fosters healthy debates, well-informed discussions, and positive interactions among netizens, as illustrated in Figure 1.1 [4].

Within this complex landscape, the imperative to mine and analyze online public opinion topics takes center stage. The process involves identifying and comprehensively understanding pivotal themes and subjects dominating online discussions. Such endeavors provide invaluable insights into the intricate dynamics of public sentiment. Equally important is assessing netizens' emotional polarity and evolution—those actively participating in online discourse. A profound comprehension of the emotional tenor of these discussions proves instrumental in accurately appraising public sentiment and prognosticating potential shifts and emerging trends.

2. Literature Review. Topic Detection and Tracking (TDT) initially emerged to uncover latent topics and monitor their developmental trajectories. The TDT model leverages mathematical statistics to condense the dimensionality of text data. However, it grapples with a limitation—its inability to harness the temporal dimension inherent in the original corpus fully. This temporal aspect is crucial for clarifying the semantic evolution of text topics over time [1].

The Online Latent Dirichlet Allocation (OLDA) model introduces a novel perspective. It posits that topic distributions exhibit both inheritance and continuity. Under this framework, the topic distribution within a historical event window serves as prior knowledge for understanding the topic's state in the current time slice. This dynamic approach allows for the online tracking of topic evolution. While the OLDA model successfully addresses the challenge of online topic modelling when new data is introduced, it grapples with a distinct issue—the redundancy stemming from intermingling old and new topics [6].

The consequence of the topic distribution redundancy is reducing the precision of topic detection and evolution analysis. It introduces a level of noise that can obfuscate the accurate identification and tracking of evolving topics within a corpus. Consequently, there is a need for more sophisticated models that can effectively disentangle the intricate interplay of old and new topics, thereby enhancing the accuracy and fidelity of topic detection and evolution analysis. A comprehensive exploration of auditory and visual environmental characteristics is embarked on a data-driven investigation encompassing 17 historical cities and towns across China. Their pioneering study harnessed posts featuring soundscapes-related keywords and street-view photographs from a prominent Chinese social media platform. The research unveiled intriguing insights into the acoustic

ambience of historic districts, revealing a symphony of artificial sounds generated by folk activities and street vendors intermingled with the soothing backdrop of natural sounds—ranging from the gentle flow of water to the melodic chorus of birdsong [19].

A novel and robust streaming media clustering algorithm is introduced, rooted in multi-edge computing, tailored explicitly for detecting streaming media traffic events. Their innovative approach incorporated domain-specific traffic-related knowledge and information. They extracted diverse elements from social media textual content to construct a heterogeneous information network (HIN) centered on traffic events. By leveraging meta-path weight calculations, the research team quantified event similarity across social media texts, offering a data-driven solution for identifying and analyzing traffic-related incidents [8].

The authors ventured into automatic emotion recognition, proposing an approach considering group types and emotional models. Their investigation delved into the facets of general datasets applicable to various emotion patterns, the overarching methodologies employed, and the reported performance metrics. The paper not only elucidated these critical properties but also analyzed the potential applications and implications of the methods discussed, charting new avenues for emotion recognition research and implementation [17].

In topic detection, the integration of emotion analysis represents a crucial dimension for unearthing hidden emotional fluctuations within topics. This interdisciplinary exploration has garnered significant attention from scholars, leading to profound advancements in two principal avenues: Firstly, researchers have sought to enhance topic detection by incorporating sentiment analysis parameters into the Latent Dirichlet Allocation (LDA) model. This approach marries the realms of topic identification and sentiment analysis, giving rise to models such as the Joint Sentiment Model (JST), Weakly Supervised Sentiment-topic Model (WSSM), and Subtopic Sentiment Combining Model (SSCM). These models share a common objective: constructing hybrid models blending topics and emotions effortlessly. By extending text-dependent topic mining to topic-dependent topic-emotion mining, these models enable the analysis of emotional evolution within topics via calculations of topic-emotion similarity. However, a drawback of these models lies in their reliance on static topic parameters, which are rooted in subjective judgments and lack dynamic adjustment capabilities, limiting their ability to track the dynamic evolution of emotions [21].

A different approach involves the integration of emotional parameters into the Online Latent Dirichlet Allocation (OLDA) model. Here, the emotional posterior of a previous time slice (t-1) is employed as the emotional prior for the subsequent time slice (t), allowing for the dynamic construction of emotional distributions across different time segments. Models such as the Time-based Subtopic and Sentiment-topic Combining Model (TSSCM) and Joint Multi-grain Topic Sentiment Model (JMTS) have emerged from this paradigm. While these models excel at dynamic topic identification, they overlook the influence of affective genetic strength in different time segments on the distribution of topic-related emotions. Furthermore, they struggle to effectively address the challenge posed by the iterative nature of emotions stemming from the interplay of old and new topics [2, 7].

In addressing the challenges above, the authors present a novel approach building upon the OLDA model, wherein emotion intensity is introduced as a pivotal element. Furthermore, it introduces the concept of emotion iteration and formulates an innovative online topic emotion recognition model. This model leverages Bayesian techniques to dynamically determine the number of topics while also harnessing the heritability of the emotion intensity operator to construct a topic-emotion distribution that unfolds over time. Through the creation of an emotion evolution channel, it adeptly discerns and tracks topic sentiment trends across diverse textual content [3, 18].

3. Research Methods.

3.1. Latent dirichlet allocation (LDA) model. LDA model constructs a three-layer Bayesian probability network by introducing Dirichlet prior parameters, an unsupervised learning model that can generate text-hidden topics. The model describes the three-layer structure relationship of document, topic and word: multiple topics have probability dependence in each document, and each topic is collected in a certain inscription according to probability. Therefore, a topic is a multinomial distribution on a thesaurus. The mixture of multiple topics constitutes the document, and the mixture of multiple feature words constitutes the topic. Assume that document d contains M documents, K topics, and V vocabulary sets. The user needs to set parameters α and β. By constructing document-topic distribution $\theta_m \left(\theta = \left\{{\theta_m}^d \mid d \in D\right\}\right)$ and topic-word distribution $\varphi_k \left(\varphi = \{\varphi_k \mid k \in [1, k]\}\right)$, the nth word $w_{m,n}$ of the mth document in the document library is obtained. The

core process of the LDA model is described as follows: (1) Set parameter α and generate document-topic distribution θ_m, namely, $\theta_m \sim \mathrm{Dir}(\alpha)$ by sampling; (2) Generate topic $Z_{m,n}$, the nth word of document m by sampling from θ_m distribution, namely, $Z_{m,n} \sim \mathrm{Mult}\ (\theta_m)$ (3). Set parameter β and generate topic $Z_{m,n}$ and its corresponding topic-word distribution φ_k by sampling, namely $\varphi_k \sim \mathrm{Dir}(\beta)$; (4) Generate the nth word $w_{m,n}$ of document m by sampling from φ_k distribution, namely, $w_{m,n} \sim \mathrm{Mult}\,(\varphi_{m,n})$.

3.2. Online latent dirichlet allocation (OLDA) model. OLDA model introduces time granularity based on the LDA model to detect the difference and continuity of online topics. By dynamically adjusting the size of the time slice, the coarse-grained requirements of the users in the time dimension of corpus analysis can be met. When processing streaming text data, the model considers that the prior and posterior probability maintains the continuity between topics, that is, the posterior weight of the word distribution φ_{t-1} in time slice $t-1$ is the prior of the word distribution φ_t in time slice t, and the Dirichlet before the word distribution φ_{t-1} in time slice t satisfies Equation (3.1) as,

$$\mathrm{Multi}\left(\varphi_k^t\right) \sim \mathrm{Dir}\left(\beta_k^t\right) \sim \mathrm{Dir}\left(\eta^\delta B_k^{t-1}\right) \tag{3.1}$$

η^δ is the heritability of the topic-word distribution on the first δ time slices. The occurrence times of words about a certain topic in the first δ time window $\{t-\delta-1, t-1\}$ are counted, and the topic-word evolution transition matrix B^{t-1}, is constructed based on this. Similarly, the document-topic distribution 0 also satisfies the similar property and the document-topic evolution transition matrix A^{t-1}, is constructed. The incremental Gibbs sampling algorithm is used in the OLDA model to approximate the over parameters α and β. Its core idea is to obtain the sample space closest to the probability distribution value by constructing a Markov chain with the probability of harvest. This algorithm considers the information of t time slice as only related to $t-1$ time slice and ignores the influence of other time slices on the corpus information. In other words, the topic-word distribution is regarded as a Markov-like chain model, and the posterior probability calculation formula of the time t slice is given by

$$P_t\left(z_i = j \mid z_{-i}, w_i\right) = \frac{\left(n_{-i,j}^{(w_i)}\right)^t + w\left(n_{-i,j}^{(w_i)}\right)^{t-1} + \beta}{\left(n_{-i,j}^{(*)}\right)^t + w\left(n_{-i,j}^{(*)}\right)^{t-1} + V\beta} \cdot \frac{\left(n_{-i,j}^{(d_i)}\right)^t + \alpha}{\left(n_{-i,*}^{(d_i)}\right)^t + K\alpha} \tag{3.2}$$

$\left(n_{-i,j}^{(d_i)}\right)^t$ represents the number of words assigned to topic j by word collection $\{V-i\}$ in the document d_i; $\left(n_{-i,*}^{(d_i)}\right)^t$ represents the total number of times the document d_i is assigned after excluding the word i; $w\left(n_{-i,j}^{(w_i)}\right)^{t-1}$ represents the number of words assigned to topic j in the $t-1$ time slice and is the same as w_i; $w\left(n_{-i,j}^{(*)}\right)^{t-1}$ represents the number of words assigned to topic j in the $t-1$ time slice. w represents heritability; V and K, represent the number of words in the corpus and the number of set topics; α and β are Dirichlet prior parameters, and the Equations give calculations.

$$\alpha_m^t = \theta_m^{t-1} \cdot A_m^{t-1} \tag{3.3}$$

$$\beta_k^t = \varphi_k^{t-1} \cdot B_k^{t-1} \tag{3.4}$$

A_m^{t-1} represents a $K \times |D^t|$-dimensional matrix arranged by θ_m^{t-1} columns of document-topic distribution on time slice $t-1$; similarly, B_k^{t-1} represents a $|V^n| \times KA_m^{t-1}$-dimensional matrix arranged by φ_k^{t-1} columns of topic-word distribution on time slice $t-1$.

3.3. Topic emotion modelling. In emotion analysis, the typical procedure involves the extraction of emotion-inducing keywords from the text, facilitated by using an emotion dictionary, followed by calculating emotion polarity. This process can generally be divided into two key stages: First, establishing a comprehensive set of emotional words is undertaken. Second, semantic proximity between emotion feature words is assessed, often employing techniques such as Similarity Calculation or Bootstrapping algorithms to derive an emotion semantic model [13, 14].

Mutual Information (MI) is widely used to calculate the semantic distance similarity between item pairs, (i_1, i_2). Generally, the emotional value of emotional feature words can be obtained by calculating its feature weight value, and the assignment of feature weight depends on the similarity distance between the combination of item vocabularies. In the index stage, all the emotion words in the text set are obtained by calculating the initial frequency value of the mixed document. The time complexity is $O\left(\left|d^2\right|\right)$ (where d is the document size) and the space complexity is $O\left(\left|T^2\right|\right)$ (where T is the number of items). The mutual information calculation is expressed as,

$$\mathrm{MI}\left(t_i, t_j\right) = \sum_{t_i \in \{0,1 | t_j \in \{0,1\}\}} \sum_{p} \log\left[\frac{p\left(t_i, t_j\right)}{p\left(t_i\right) p\left(t_j\right)}\right] \times p\left(t_i, t_j\right) \tag{3.5}$$

where $p\left(t_i, t_j\right)$ represents the probability of t_i and t_j co-appearing in the same document; $p\left(t_i\right)$ and $p\left(t_j\right)$ represent the probability that the document contains t_i and t_j, respectively. The above probability can be obtained from the initial corpus using maximum likelihood estimation.

The emotional polarity in the emotional dictionary has three kinds: positive, negative, and neutral. The classification of affective tendency generally sets an affective threshold and compares the mutual information of the calculated new item i_1 with that of the known effective item i_2. If $\mathrm{MI}\left(i_1, i_2\right)$ is less than the threshold, (i_1, i_2) is put into different affective tendencies [12, 11].

Emotional dictionaries often contain emotional, turning conjunctions and negative words. It is necessary to use the conditional clause after the turning conjunctions to replace the whole sentence before calculating the feature weight value of new words (emotional and negative words). The multi-feature linear fusion method calculates the comprehensive affective value, effectively avoiding the uncertainty of affective inclination. The calculation formula is:

$$\mathrm{SO}_{\mathrm{neg}}\left(i_{\mathrm{new}}\right) = \frac{\sum_{i_{\mathrm{pos}} \in \{0,m\} i_{\mathrm{pos}} \in I_{\mathrm{pos}}} \mathrm{MI}\left(i_{\mathrm{new}}, i_{\mathrm{pos}}\right) \cdot \mathrm{neg}}{\left|I_{\mathrm{pos}}\right|} + \frac{\sum_{i_{\mathrm{agg}} \in \{0,n\} i_{\mathrm{neg}} \in I_{\mathrm{neg}}} \mathrm{MI}\left(i_{\mathrm{new}}, i_{\mathrm{neg}}\right) \cdot \mathrm{neg}}{\left|I_{\mathrm{neg}}\right|} \tag{3.6}$$

where i_{new} represents a new word; i_{pos} and i_{neg} respectively represent the classified positive and negative emotion words in the old words. neg is the negative sign of the sentence where the nord is found. When $neg = 1$, it indicates that there is no negative word on the right of the new word, and the affective tendency is consistent with the affective feature words in front. When $neg = -1$, it indicates negative words on the right side of new words, and the affective tendency is opposite to the affective feature words in front. I_{pos} and I_{neg} represent the positive emotion word set and negative emotion word set of the emotion word list, respectively; m and n represent the number of words in positive and negative emotion word sets, respectively. When $\mathrm{SO}_{\mathrm{neg}}\left(i_{\mathrm{new}}\right) > 0$, it indicates that i_{new} has a positive tendency and is added to the positive emotion word set. When $\mathrm{SO}_{\mathrm{neg}}\left(i_{\mathrm{new}}\right) < 0$, it indicates that there is a negative tendency, and a negative emotion word set is added. When $\mathrm{SO}_{\mathrm{neg}}\left(i_{\mathrm{new}}\right) = 0$, there is no emotional inclination [10, 5].

3.4. OTSRM model description. The OTSRM model is a three-layer Bayesian network. Let the focus word N^x of t time slice contains K^t feature words and M^t emotion words. Firstly, the model obtains β priori of $t-1$ time slice according to the focus word distribution θ_m^{t-1} of $t-1$ time slice. According to the focus words obtained, the feature word $w_{m,n}^{t-1}$ is selected from feature word distribution $\emptyset_k^{t-1}$ then emotion word $S_{m,n}^{t-1}$ is extracted based on emotion word distribution μ_k^{t-1} and emotion intensity is calculated. Finally, topic emotion words with maximum emotion value are obtained through topic emotion calculation.

3.5. Calculation of emotional intensity. The OLDA model maintains the continuity between topics through topic heritability w. It is assumed that the variable distribution of t-time slice is only affected by $t-1$ time slice and has nothing to do with the text information in the previous time slice so that it can be regarded as a topic inheritance. OTSRM model refers to the property of topic inheritance and introduces the emotion iteration thought into emotion analysis. The topics-emotion distribution μ_m^t of t-time slice is regarded as the posterior of μ_m^t in $t-1$ time slice. By calculating the heritability λ of emotion intensity of $t-1$ time slice, the topic emotion intensity of t-time slice is obtained.

Emotional intensity is similar to topic intensity, which can dynamically measure the stability of topic emotion in the time dimension. Documents describing a topic have high probability distribution values on a

few topics, and sentiment words describing a topic will also show relatively high probability distribution values on one or a few topics. Similarly, suppose the probability distribution of a topic in each emotion word is relatively average. In that case, it can be determined that the emotion expressed by the document is relatively balanced and has no clear emotional tendency. In this article, Shannon's information entropy is used to represent the emotional concentration degree of the topic. The normalized topic of emotional weight w_m^t is calculated using Equations (3.7) and (3.8):

$$E\left(z_m^t\right) = -\sum_{m=1}^{M}\sum_{k=1}^{K}\mu_{m,k}^{t-1}\log_2\mu_{m,k}^{t-1} \tag{3.7}$$

$$W_m^t = 1 - \frac{E\left(z_m^t\right) - \min\left\{E\left(z_1^t\right),\cdots,E\left(z_M^t\right)\right\}}{\max\left\{E\left(z_1^t\right),\cdots,E\left(z_M^t\right)\right\} - \min\left\{E\left(z_1^t\right),\cdots,E\left(z_M^t\right)\right\}} \tag{3.8}$$

$\mu_{m,k}^{t-1}$ represents the emotion distribution of document m under the k th topic under the $t-1$ time slice. When the topic belongs to only one emotion word, the emotion information quotient under the topic of the document is 0 , and the emotion weight W_m^t is 1. In particular, when W_m^t is 0, the topic is evenly distributed on K emotions, then the corresponding affective information entropy is maximum, indicating that the affective distribution of this topic is relatively broad and does not contribute to the affective of the topic. The Formula for calculating the hyperparameter γ_m^t of topic emotion distribution in time slice t is:

$$\gamma_m^t = \lambda_m^{t-1}\cdot R_m^{t-1} \tag{3.9}$$

where R_m^{t-1} represents the topic emotion matrix in $t-1$ time slice; λ_m^{t-1} represents the heritability of topic emotion in $t-1$ time slice. The Formula for calculating the heritability λ_m^t of topic emotion in t time slice is

$$\lambda_m^t = \frac{1}{K^t}\left(K^t - \mathrm{rank}_k^{t-1}\right) \tag{3.10}$$

After calculating the emotional weight of each topic, the Formula for calculating the emotional intensity $J\left(z_k^t\right)$ of the topic is

$$J\left(z_k^t\right) = \frac{\sum_{m=1}^{M} W_m^{t-1}\mu_{m,k}^{t-1}}{M} \tag{3.11}$$

where $\mu_{m,k}^{t-1}$ represents the emotion distribution of document m under the k th topic under the $t-1$ time slice; W_m^{t-1} represents topic emotion weight under $t-1$ time slice; M is the number of documents.

3.6. Iterative model solution. OTSRM model takes dominant variable time t and word $w_{m,n}$ as initial input values, topic variable $\boldsymbol{Z}$ and emotion variable s as implicit variables, and uses the improved Gibbs sampling algorithm to solve the joint a posteriori probability function of topic distribution θ, feature word distribution φ and emotion distribution μ and the calculation formula is

$$P_t\left(z_i = k \mid z_{-i}, S_{-i}, w_i\right) = \frac{\left(n_{j,s}^{(d_i)}\right)^t + \alpha^t}{\left(n_{*,s}^{(d_i)}\right)^t + K^t\alpha^t}\cdot\frac{\left(n_{j,s}^{(w_j)}\right)^t + w\left(n_j^{(*)}\right)^{t-1} + \beta^t}{\left(n_{j,s}^{(*)}\right)^t + w\left(n^{(*)}\right)^{t-1} + V^t\beta^t}\cdot\frac{\left(n_{j,s}^{(w_i)}\right)^t + w\left(n_{j,s}^{(w_i)}\right)^{t-1} + \gamma^t}{\left(n_{j,s}^{(*)}\right)^t + w\left(n_{j,s}^{(*)}\right)^{t-1} + L^t\gamma^t} \tag{3.12}$$

where, $\left(n_{j,s}^{(d_i)}\right)^t$ represents the number of words assigned to emotion s by feature word j in the document d_i of time slice t; $\left(n_{*,s}^{(d_i)}\right)^t$ represents the total number of words that feature word j assigned to emotion s in the document d_i of time slice t; $\left(n_{j,s}^{(w_j)}\right)^t$ represents the number of words assigned by t time slice word w to feature word j and emotion word j; $\left(n_{j,s}^{(*)}\right)^t$ represents the total number of words assigned by t time slice word w to feature word j and emotion word s; $w\left(n_{j,s}{}^{(w_i)}\right)^{t-1}$ represents the heritability of emotion s assigned by $t-1$ time

slice word w to feature word j; $w\left(n_{j,s}^{(*)}\right)^{t-1}$ represents the sum of heritability of emotion s assigned by $t-1$ time slice word w to feature word j. Therefore, in each sampling, Equation (3.12) is iterated until relatively stable θ, φ, and μ distributions are obtained. Equation (3.15) of Formula (3.13) shows the corresponding probability distribution update.

$$\theta_{(z=j,s_i)}^{(d_i)} = \frac{\left(n_{j,s}^{(d_i)}\right)^t + \alpha^t}{\left(n_{*,j}^{(d_i)}\right)^t + K^t\alpha^t} \tag{3.13}$$

$$\varphi_{(z=j,s_i)}^{(d_i)} = \frac{\left(n_{j,s}^{(w_i)}\right)^t + w\left(n_j^{(*)}\right)^{t-1} + \beta^t}{\left(n_{j,s}^{(*)}\right)^t + w\left(n_j^{(*)}\right)^{t-1} + V^t\beta^t} \tag{3.14}$$

$$\mu_{(z=j,s_i)}^{(d_i)} = \frac{\left(n_{j,s}^{(w_i)}\right)^t + w\left(n_{j,s}^{(w_i)}\right)^{t-1} + \gamma^t}{\left(n_{j,s}^{(*)}\right)^t + w\left(n_{j,s}^{(*)}\right)^{t-1} + L^t\gamma^t} \tag{3.15}$$

3.7. Topic emotion calculation.

3.7.1. Topic similarity calculation. A common way to measure the similarity between two probability distributions is the KL distance. KL distance can also calculate the topic-word distribution difference in adjacent time slices. However, the asymmetry of KL distance cannot solve the symmetric topic distribution function. This article introduces relative entropy into topic similarity measurement, and a topic similarity calculation based on relative entropy is established. The specific Formula is as follows:

$$\begin{aligned}\mathrm{Sim}_{\mathrm{KL}}\left(\varphi_k^{t-1}, \varphi_k^t\right) &= -\frac{1}{2}\left[\mathrm{KL}\left(\varphi_k^{t-1}, \varphi_k^t\right) + \mathrm{KL}\left(\varphi_k^t, \varphi_k^{t-1}\right)\right] \\ &= \frac{1}{2}\left[\sum_{w\in V, k\in K} p(w)\log\frac{\varphi_k^{t-1}}{\varphi_k^t} + \sum_{w\in V, k\in K} q(w)\log\frac{\varphi_k^t}{\varphi_k^{t-1}}\right]\end{aligned} \tag{3.16}$$

$p(w)$ and $q(w)$ represent the probability of feature word w appearing in the distribution of topic Z' and Z^{t-1}, respectively.

3.7.2. Calculation of topic affective similarity. OTSRM model establishes the topic emotion model, which depends on the set of emotion words by calculating the maximum emotion value of emotion words under different topics. The specific calculation method is as follows: firstly, the position of the sentence in which the emotional word w_i of topic z is located and determined, whether there are turning conjunctions and negative words in the sentence in which it is located are determined, and the value of the negative flag neg is obtained, and the result is used to calculate the emotion $S(w_i)$, specifically, as follows:

$$S(w_i) = \mathrm{SO}_{\mathrm{neg}}(w_i) \cdot L(w_i) \tag{3.17}$$

where $L(w_i)$ represents the value of modifying degree words; $\mathrm{SO}_{\mathrm{neg}}\ (w_i)$ represents the comprehensive affective value of w_i. Finally, the expectation of all emotion words under topic $\boldsymbol{z}$ is taken as its final emotion value, and the calculation formula is:

$$S(z) = \frac{\sum_{i=1}^n S(w_i)}{n} \tag{3.18}$$

3.8. Experimental data. The experimental data in this article come from the news. The GooSeeker web crawler software screened five network hot events as empirical analysis cases.

Due to the space limitation, this article only takes Datal as an example to illustrate the process of topic emotion evolution analysis of OTSRM. First, the Chinese word segmentation software NLPIR cuts the text

Table 4.1: Results of feature word identification on the topic of "murder case in a mountain"

Topic	Feature words and their probabilities	Report time and probability
1	Mountain 0.013/ Traffic 0.011/ cut 0.009/ BMW male 0.012/ Dispute 0.016/ Electric vehicle 0.007/ death 0.017/ chase 0.019/ injury 0.013/ pick up	$t_1 = 0.218$, $t_2 = 0.257$, $t_3 = 0.201$, $t_4 = 0.233$, $t_5 = 0.246$, $t_6 = 0.225$
2	Procuratorate 0.021/ filing, 0.022/ conflict, 0.022/ alarm, 0.022/ alarm, 0.022/ first aid, 0.022/ drunk, 0.022/Harm 0.022/ Control 0.022/ suspected 0.020	$t_2 = 0.235$, $t_3 = 0.271$, $t_4 = 0.203$, $t_5 = 0.239$, $t_6 = 0.225$
3	Tattoo 0.021/ criminal record 0.019/ gang-related 0.021/ Good Samaritanism 0.021/ pawn shop 0.021/ burden 0.020/ illness 0.019/ Difficulties 0.019/ loans 0.019/ Cheonan Society 0.019	$t_2 = 0.243$, $t_3 = 0.296$, $t_4 = 0.233$, $t_5 = 0.277$
4	Legitimate 0.032/ safety 0.032/ excessive 0.032/ difference 0.032/ dodge 0.032/ Run 0.032/ calm 0.032/ law 0.032/ Injury 0.032/ Subjective 0.032	$t_3 = 0.312$, $t_4 = 0.217$, $t_5 = 0.296$, $t_6 = 0.211$
5	Notification 0.032/ justification 0.032/ defense 0.032/ Revocation 0.032/ Punishment 0.032/ punishment 0.032/ innocence 0.032/ immunity 0.032/ Determination 0.032/ compliance 0.032	$t_5 = 0.326$, $t_6 = 0.203$

Note: t_i indicates that the event occurred on the ith day.

into word sets and filters the stop words. Then, word sets corresponding to the report text d_c^t and its comment d_j^t belonging to the same time slice are combined into mixed word sets D_c^t. The emotion word frequency in d_j^t was measured by $(t = 1)$ day after the occurrence of the event $t = 1$ for 6 consecutive days. Finally, the word sets of t time slice is modelled successively, and an online sentiment dictionary is established according to the OTSRM model [15, 22].

4. Results and Discussion. In the OTSRM model, when $t = 1$ timepiece, the initial values of $\alpha^t, \beta^t, \gamma^t$ were set as 0.5, 0.1 and 0.1, respectively, topic feature word W_c. and emotion word W_s were set as 15, the probability generation threshold. and emotion threshold of topic features were both 0.2 and Gibbs sampling was set 2000 times.

4.1. Topic identification. The topic feature words identified by the OTSRM model and their corresponding probability distribution are shown in Table 4.1. Each topic is represented by the top 10 feature words with high probability.

From Table 4.1, a mountain murder event contains five topics. Topic 1: A mountain death caused by a traffic dispute, the duration of which is $t_1 \sim t_6$; Topic 2: The procuratorial organ files a case for investigation, the duration is $t_2 \sim t_6$; Topic 3: Background report of both parties involved, duration is $t_2 \sim t_5$; Topic 4: How to determine self-defense, duration is $t_3 \sim t_6$; Topic 5: The police decided the case. The duration is $t_5 \sim t_6$. According to Table 4.1, the evolution process of all topics in different time slices can be obtained (as shown in Figure 4.1).

From Figure 4.1, different topics coexist in the time dimension. For example, topic 2 and topic 3 overlap in the $t_2 \sim t_5$ time slice that is, "investigation by the procuratorial organ" and "background reports of both parties involved" coexist in multiple news reports, reflecting the diversity of perspectives of news reports.

4.2. Analysis of emotional evolution. According to the topic-emotion distribution, the emotional intensity of $t_1 \sim t_6$ time slice was calculated, and the topic-emotion matrix was obtained. At the same time, topic emotion words were selected, and the topic emotion values of different time slices were calculated using Equation (3.18). Taking topic 1 as an example, Table 4.2 shows the emotional outcomes identified. Due to the space limitation, other topic identification results are similar to Topic 1, which will not be repeated here. To more clearly show the emotion evolution recognition process of the OTSRM model on Data1, the emotion

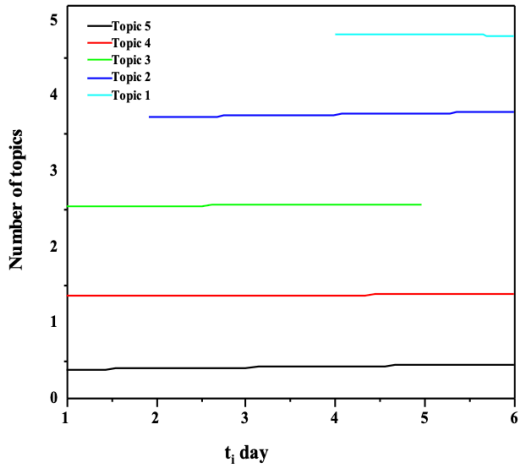

Fig. 4.1: Topic evolution process

Table 4.2: Emotion recognition results of topic 1 in $t_1 \sim t_6$ time slices

Time	Emotional words and their probabilities	Affective value
t_1	Scum 0.015/ Deserved 0.017/ innocent 0.021/ gas relief 0.018/ death 0.026/ Severe punishment 0.017/ impulse 0.026/ Courage 0.013/ show off 0.011/ brute 0.015-0.78	-0.78
t_2	Harm 0.014/ release 0.022/ Thorough investigation 0.013/ Quick 0.016/ Anger 0.017/ Compensation 0.014/ rampant 0.018/ light sentence 0.015/ Expectation 0.018/ innocence 0.026-0.42	-0.42
t_3	Hit 0.016/ disappointed 0.017/ hateful 0.021/ dead 0.016/ pathetic 0.017/ pity 0.018/ release 0.019/ Pity 0.022/ heavy judgment 0.017/ arrogance 0.018-0.26	-0.26
t_4	Bully 0.031/ Helpless 0.029/ Support 0.036/ Top 0.021/ innocent 0.032/ scum 0.024/ excessive 0.024/ sad 0.027/ harm 0.029/ sit and wait 0.033-0.34	-0.34
t_5	Kill 0.031/ legitimate 0.036/ helpless 0.027/ hateful 0.027/ poor 0.032/ Judgment 0.028/ gas relief 0.034/ sympathy 0.031/ hateful 0.026/ eradicate 0.032 0.16	0.16
t_6	Hope 0.027/ Support 0.034/ Like 0.029/ Top 0.035/ positive 0.028/ Fair 0.032/ justice 0.027/ stand 0.029/ happy 0.029/ Hold 0.027 0.43	0.43

values of five topics under different time slices were calculated respectively, and the dynamic comment emotion information evolution diagram was obtained, as shown in Figure 4.2.

From Figure 4.2, topic emotion presents a dynamic evolution over time. Topic 1 showed great emotional fluctuation, which was caused by the fact that in the initial stage of the event, netizens concentrated on expressing their condemnation and anger towards "BMW Man", which reflected strong negative emotion. When t=4, the topic of public concern gradually turned to the discussion on the judgment conditions of justifiable defence, and they expressed their sympathy for Haiming. The positive emotion they showed offset part of the negative emotion, so the overall emotion value was low-intensity negative emotion. When t=6, as the police determined the case as a justifiable defence, the positive emotion of the public reached the highest value, and strong positive emotion appeared.

The topic emotion of topic 2 and topic 3 is near the neutral emotion, which reflects the tangled and confused mentality of the public, that is, there is confusion about the defining conditions of justifiable defence and excessive defence, and the public is guessing the conclusion without clear emotional tendency. Topic 4 showed relatively stable positive emotions, and the public's sympathy for the weak party in a justifiable defence dominated. Topic 5 shows strong and stable positive sentiment, and the public's sentiment is similar to the

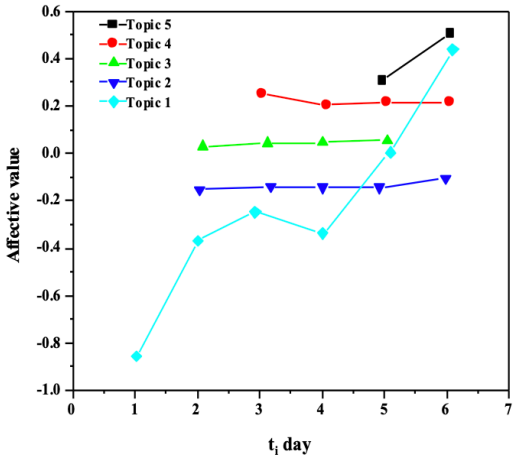

Fig. 4.2: Evolution distribution of topic emotion

Table 4.3: Comparison of emotion recognition accuracy

Model	Data 1	Data 2	Data 3	Data 4	Data 5
JST	66.16	67.09	69.81	71.36	72.95
WSSM	68.82	72.64	65.38	74.91	73.66
SSCM	74.46	75.92	72.65	75.64	74.68
TSSCM	77.68	78.84	76.31	78.88	79.35
OTSRM	78.94	80.13	79.96	81.03	80.62

sentiment tendency of topic 1 in the later stage, both of which express firm support for the judicial department to maintain social justice.

4.3. Model evaluation. In order to verify the algorithm performance of the OTSRM model, data sets of other 4 events were substituted into the model according to the above experimental process, and similar experimental results were obtained. In addition, this article uses accuracy rate, recall rate and F value as three indicators to make a comprehensive evaluation compared with the algorithms provided in the literature. Where, the initial values of the superparameter α^t, β^t and γ^t of the above model are set as 0.5,0.1 and 0.1, respectively. Topic feature word W_c and emotion word W_s were set as 15. The probability generation and emotion threshold of topic features were 0.2, and Gibbs was set to sample 1000 times. The accuracy and recall rates of the five models on the five data sets are shown in Table 4.3 and Table 4.4, respectively, and the comparison results of the F value are shown in Figure 4.3.

As can be seen from Figure 4.3, the highest emotion recognition accuracy rates of OTSRM are 85.56% and 81.03%. TSSCM and OTSRM models are significantly better than JST, WSSM and SSCM models in emotion classification because the algorithm based on the OLDA model comprehensively considers the topic relevance of different time slices. The number of topics is set dynamically according to probabilistic topics to achieve the purpose of topic temporal and spatial modelling. However, the prior topic emotion parameters of JST, WSSM and SSCM models are greatly affected by subjective experience, which easily causes skew of emotion mining, thus reducing classification accuracy. Although the TSSCM model can dynamically identify topics, it does not consider the transitivity of the topic's emotional intensity. It cannot effectively solve the problem of emotional iteration caused by mixing old and new topics. The algorithm in this article makes up for the shortcomings of the above algorithms. By introducing the emotion intensity with heritability, the topic in different time slices is acquired dynamically, and the mixed model of the coexistence of topic and emotion is established, effectively

Table 4.4: Comparison of emotion recognition recall rate

Model	Data 1	Data 2	Data 3	Data 4	Data 5
JST	76.35	74.32	77.45	78.62	76.64
WSSM	77.92	78.94	74.49	79.78	78.26
SSCM	79.81	80.18	79.46	80.64	79.98
TSSCM	81.67	83.94	82.16	83.37	82.21
OTSRM	83.33	84.62	82.77	85.56	84.33

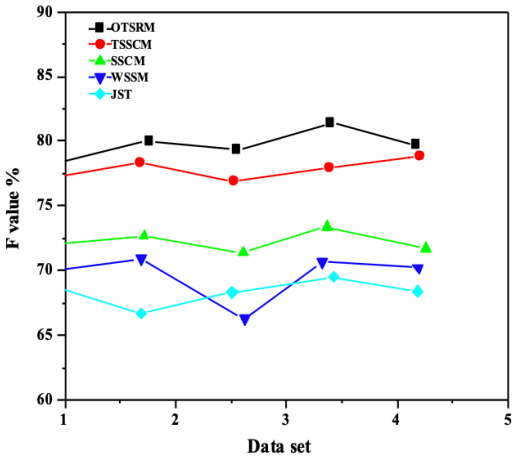

Fig. 4.3: Comparison results of F-value of emotion recognition

improving the recognition accuracy of topic and emotion.

5. Conclusion. An innovative approach for Korean topic sentiment analysis is introduced in this paper to leverage social media big data clustering. The basis of the proposed contribution is developing the Online Topic Sentiment Recognition Model (OTSRM), which introduces several key elements to enrich the topic sentiment analysis in Korean social media discussions. The OTSRM model is characterized by constructing an emotion intensity operator, establishing an emotion evolution channel based on posterior probability, and acquiring feature word and emotion word distribution matrices. These components work synergistically to enable a dynamic representation of the evolving emotional recognition within textual content. Using the relative entropy method, we calculate the maximum emotion value associated with the current topic focus, offering valuable insights into the real-time evolution of topic emotions within the text. The experimental validation has underscored the OTSRM model's effectiveness in topic emotion evolution analysis. It has demonstrated robust performance and the ability to capture the tones of emotional fluctuations within Korean social media discourse. These findings confirm the relevance and utility of the suggested model in uncovering the dynamic emotional undercurrents that shape public sentiment within online discussions.

Acknowledgements. The Key Research Project supported by Social Science research Fund of Lingnan Normal University in 2022: A Study on the Image of China in Foreign Novels in the Perspective of Powerful Cultural Country [WZ2204].

REFERENCES

[1] J. Allan, *Introduction to topic detection and tracking*, in Topic Detection and Tracking: Event-based Information Organization, Springer, 2002, pp. 1–16.

[2] Y. Alotaibi, M. N. Malik, H. H. Khan, A. Batool, A. Alsufyani, S. Alghamdi, et al., *Suggestion mining from opinionated text of big social media data.*, Computers, Materials & Continua, 68 (2021).
[3] M. Asgari-Chenaghlu, M.-R. Feizi-Derakhshi, L. Farzinvash, M.-A. Balafar, and C. Motamed, *Topic detection and tracking techniques on twitter: a systematic review*, Complexity, 2021 (2021), pp. 1–15.
[4] A. Barradas, A. Tejeda-Gil, and R.-M. Cantón-Croda, *Real-time big data architecture for processing cryptocurrency and social media data: A clustering approach based on k-means*, Algorithms, 15 (2022), p. 140.
[5] Y. Deng and D. Liu, *A multi-dimensional comparison of the effectiveness and efficiency of association measures in collocation extraction*, International Journal of Corpus Linguistics, 27 (2022), pp. 191–219.
[6] Y. Guo, F. Wang, C. Xing, and X. Lu, *Mining multi-brand characteristics from online reviews for competitive analysis: A brand joint model using latent dirichlet allocation*, Electronic Commerce Research and Applications, 53 (2022), p. 101141.
[7] D. Hyun and S. Lee, *A study of big data analysis regarding smartphone user satisfaction: Utilizing sentiment analysis based on social media data*, Korean Journal of Converging Humanities, 9 (2021), pp. 7–35.
[8] Y. Ji, J. Wang, Y. Niu, and H. Ma, *Reliable event detection via multiple edge computing on streaming traffic social data*, IEEE Access, (2021).
[9] Y. Kim, D. Sohn, and S. M. Choi, *Cultural difference in motivations for using social network sites: A comparative study of american and korean college students*, Computers in Human Behavior, 27 (2011), pp. 365–372.
[10] X. Lv and M. Li, *Application and research of the intelligent management system based on internet of things technology in the era of big data*, Mobile Information Systems, 2021 (2021), pp. 1–6.
[11] J. Meng and R. Tench, *Strategic communication and the global pandemic: Leading through unprecedented times*, International Journal of Strategic Communication, 16 (2022), pp. 357–363.
[12] M. Mosleh, G. Pennycook, and D. G. Rand, *Field experiments on social media*, Current Directions in Psychological Science, 31 (2022), pp. 69–75.
[13] K. R. Nastiti, A. F. Hidayatullah, and A. R. Pratama, *Discovering computer science research topic trends using latent dirichlet allocation*, Jurnal Online Informatika, 6 (2021), pp. 17–24.
[14] J. St John, K. St John, and B. Han, *Entrepreneurial crowdfunding backer motivations: a latent dirichlet allocation approach*, European Journal of Innovation Management, 25 (2022), pp. 223–241.
[15] J. Tong, L. Shi, L. Liu, J. Panneerselvam, and Z. Han, *A novel influence maximization algorithm for a competitive environment based on social media data analytics*, Big Data Mining and Analytics, 5 (2022), pp. 130–139.
[16] S. K. Uppada, K. Manasa, B. Vidhathri, R. Harini, and B. Sivaselvan, *Novel approaches to fake news and fake account detection in osns: user social engagement and visual content centric model*, Social Network Analysis and Mining, 12 (2022), p. 52.
[17] E. A. Veltmeijer, C. Gerritsen, and K. V. Hindriks, *Automatic emotion recognition for groups: a review*, IEEE Transactions on Affective Computing, 14 (2021), pp. 89–107.
[18] T. W. Wibowo, S. H. M. B. Santosa, B. Susilo, and T. H. Purwanto, *Revealing tourist hotspots in yogyakarta city based on social media data clustering*, Geo Journal of Tourism and Geosites, 34 (2021), pp. 218–225.
[19] H. Xie, Y. He, X. Wu, and Y. Lu, *Interplay between auditory and visual environments in historic districts: A big data approach based on social media*, Environment and Planning B: Urban Analytics and City Science, 49 (2022), pp. 1245–1265.
[20] C. Yang, G. Su, and J. Chen, *Using big data to enhance crisis response and disaster resilience for a smart city*, in Proceedings of the 2nd International Conference on Big Data Analysis, Beijing, China, 2017, IEEE, pp. 504–507.
[21] T. Zhou, K. Law, and D. Creighton, *A weakly-supervised graph-based joint sentiment topic model for multi-topic sentiment analysis*, Information Sciences, 609 (2022), pp. 1030–1051.
[22] Y. Zhou, L. Liao, Y. Gao, R. Wang, and H. Huang, *Topicbert: A topic-enhanced neural language model fine-tuned for sentiment classification*, IEEE Transactions on Neural Networks and Learning Systems, (2021).

Edited by: Venkatesan C
Special issue on: Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications
Received: May 13, 2023
Accepted: Oct 9, 2023