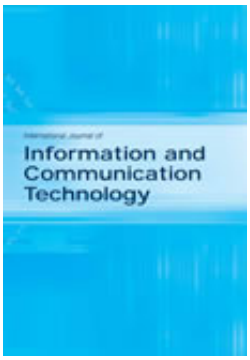
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Ideological opinion clustering identification based on Gibbs sampling in social new media environment

Lei Yang, Dongbo Xu

Article History:

Received:	31 December 2024
Last revised:	13 January 2025
Accepted:	14 January 2025
Published online:	21 March 2025

Ideological opinion clustering identification based on Gibbs sampling in social new media environment

Lei Yang

School of History,
Nanjing University,
Nanjing, 210023, China
Email: 602023100048@smail.nju.edu.cn

Dongbo Xu*

Center of National Security and National Defence Education,
Nanjing Agricultural University,
Nanjing, 210095, China
Email: lwtgyue@126.com
*Corresponding author

Abstract: The effect of ideological public opinion in social opinion is growing as social new media develops rapidly. Effective mining of crucial information from the vast social new media data has become a hot issue in present opinion analysis study. Thus, this work presents an ideological opinion analysis model, GibbsCluster, derived from the combination of Gibbs sampling and K-means clustering. Using Gibbs sampling, the model separates opinion data into groups by means of the K-means and combines with sentiment analysis for fine-grained opinion classification. In the combined effect of opinion clustering and sentiment analysis, the experimental findings reveal that the GibbsCluster model much beats conventional approaches. This work also tests the adaptability of the model in other social platforms and creates a creative evaluation approach to fully evaluate its performance by accuracy and F1-score.

Keywords: social new media; ideological opinion; Gibbs sampling; K-means clustering; sentiment analysis.

Reference to this paper should be made as follows: Yang, L. and Xu, D. (2025) 'Ideological opinion clustering identification based on Gibbs sampling in social new media environment', *Int. J. Information and Communication Technology*, Vol. 26, No. 5, pp.22–38.

Biographical notes: Lei Yang received her Master's degree in Southeast University in 2019. From 2019 to now, she works in Jiangxi Flight University. Currently, she studies in Nanjing University. Her research interests are included National security governance and National security education.

Dongbo Xu received his Doctor's degree from the Nanjing University in 2022. From 2012 to now, he works in Nanjing Agricultural University. His research interests are included national security governance and OSINT.

1 Introduction

Social new media has grown to be a major venue for information sharing in contemporary culture as social networks develop quickly (Kapoor et al., 2018). Microblogging, WeChat, Facebook, Twitter, and other platforms have evolved from tools for daily contact to means of fast spreading all kinds of knowledge and viewpoints. Particularly in the sharing of ideas and opinions, social new media's importance has grown progressively important (Linders, 2012). By means of comments, likes, retweets, and other interactions, users of these platforms not only express their own emotions, perspectives, and ideas but also shape the impressions and emotions of others, so generating a sizable and dynamic public opinion network (Loader and Mercea, 2011).

Ideological public opinion is the general feelings and ideological responses of the public to hot social concerns, political affairs, cultural events, etc. inside a given period of time. It has high emotional inclination, spreads rapidly and broadly. Thus, for governments and businesses, knowing and analysing ideological public opinion in real-time can assist identify possible social issues and hazards in time and subsequently enable the implementation of sensible remedies. Thus, ideological public opinion analysis has not only attracted great attention in the field of social sciences but also evolved into a major study focus in the field of information technology.

In this regard, one of the hot topics in research is ideological public opinion analysis in the social media environment (Xu et al., 2022). The main challenges of modern technical research are now how to effectively extract meaningful information from the enormous social new media data and how to correctly recognise the emotional tendency and transmission method of public opinion.

Text mining approaches, which typically extract keywords, subjects, and sentiment information from the text by means of word segmentation, word frequency statistics, and sentiment dictionary matching, provide the foundation of most traditional approaches of public opinion study (Mostafa, 2013). Among the particular techniques are attitude dictionaries, word frequency statistics, topic models, and so forth. Using the sentiment dictionary approach as an example, it computes the sentiment polarity of the words to evaluate the sentiment tendency of the text by matching the sentiment terms in the text with a pre-constructed sentiment lexicon. Although the technique is basic and easy to apply, its performance suffers when dealing with works with complicated and cryptic emotional expression (Traver, 2010). Furthermore constrained by the coverage of the vocabulary is the sentiment lexicon technique, which makes it challenging to deal with varied and often shifting opinion contents.

Topic modelling – using latent Dirichlet allocation (LDA) model – is another often used text mining technique (Jelodar et al., 2019). Assuming that every document comprises of a mix of several topics, LDA is a generative model that can efficiently mine the possible topic distribution in a text. LDA has certain constraints even if it can more clearly expose the topic structure of opinion texts (Sharma et al., 2022). For instance, the LDA model is challenging to manage semantic information in short texts and sensitive to hyperparameters. Furthermore, the LDA model sometimes ignores the dependencies between words and thinks that the words inside every document are independent, therefore partially failing in capturing the deeper semantics in the text.

Traditional text mining techniques progressively reveal their bottlenecks in large-scale data processing in view of the explosive rise in data volume. Machine

learning-based approaches for public opinion analysis have progressively become mainstream in order to address these challenges. By means of training models, machine learning techniques automatically learn characteristics from data and handle tasks including categorisation and prediction. Common machine learning techniques comprise support vector machine (SVM), decision tree, random forest, neural network, etc. (Bansal et al., 2022).

Machine learning techniques improve generalisation capacity of the model by optimising in a larger dimensional feature space than conventional approaches. Machine learning techniques do, nevertheless, also have certain flaws (Dobbelaere et al., 2021). First, these techniques typically need a lot of labelled data for training and for opinion data in some special domains; second, machine learning models depend on high quality of data and the interference of noisy data may cause model performance to drop.

With the fast growth of unsupervised learning and probabilistic modelling in recent years, academics have progressively suggested more flexible and effective approaches for opinion analysis. Specifically, Gibbs sampling applied in topic modelling has significantly enhanced the identification of possible patterns in large-scale text data. Concurrently, the analysis of opinion data uses clustering techniques – a classic tool for unsupervised learning – often to detect natural groups and sentiment trends in the data. These methods together offer fresh approaches for ideological opinion analysis, however how best to combine sentiment analysis with clustering is still a topic of study deserving of more investigation. This work proposes an ideological opinion analysis model, GibbsCluster, based on Gibbs sampling and K-means clustering, to precisely identify the emotional inclinations of various opinion groups by grouping social new media data and aggregating the findings of sentiment analysis.

The main innovations include:

1 Combining Gibbs sampling and clustering methods for ideological opinion analysis. For ideological opinion analysis in social new media, the GibbsCluster model presented in this work creatively blends Gibbs sampling with K-means clustering algorithm. Gibbs sampling models possible themes in text data together with clustering techniques to automatically detect the emotional tendencies of various groups on social media, therefore improving the accuracy and interpretability of opinion analysis.

2 Multi-dimensional sentiment analysis combined with clustering. This work not only presents the dimension of sentiment analysis in clustering analysis but also improves the effect of opinion clustering by means of sentiment classification. While grouping various opinion groups, the model painstakingly classifies the sentiment tendency of every group, therefore enabling a better knowledge of the public's sentiment swings during the process of opinion spread.

3 Innovative evaluation index design. This work creatively combines the evaluation indexes of sentiment analysis on the basis of conventional clustering evaluation indexes (e.g., accuracy and F1-score) so completely assessing the performance of GibbsCluster model. This work develops a complete assessment approach that satisfies the requirements of opinion analysis by fully addressing the accuracy of sentiment classification and clustering effect, so offering a new reference dimension for the performance evaluation of opinion analysis algorithms.

4 Potential for cross-domain application. This research presents a methodology that is not only relevant for ideological public opinion analysis but also for other domains including product evaluation study, social media crisis management, and public health information distribution. In these fields, the model produces more sophisticated analysis findings by correctly spotting the sentiment tendency and opinion distribution of various groups.

2 Relevant technologies

2.1 Gibbs sampling

Widely used to generate approximative inferences from complex probability distributions, Gibbs sampling is a sampling method grounded on Markov chain Monte Carlo (MCMC) approach (Gnanasekaran and Balaji, 2013). Gibbs sampling is particularly appropriate for high-dimensional data in ideological opinion analysis since it can help us to uncover possible opinion subjects from a great volume of textual data in the social new media environment.

Assume we intend to sample from a high-dimensional joint distribution $p(x)$, in which case x is a high-dimensional vector of random variables, i.e.,

$$x = \left(x_1, x_2, \ldots, x_d\right) \tag{1}$$

Gibbs sampling is fundamentally the method of progressively sampling from the target distribution by one-by-one variable update of the conditional distribution. We specifically change the value of the i^{th} variable x_i to be sampled in line with its conditional distribution $p(x_i|x_{-i})$ at the t^{th} iteration. The equation is as follows:

$$x_i^{(t+1)} \sim p\left(x_i \left| x_{-i}^{(t)}\right.\right) \tag{2}$$

where $x_{-i}^{(t)}$ indicates the other variables eliminated with the i^{th} variable. Through alternately updating every variable, Gibbs sampling progressively converges to the target distribution in every iteration.

Assume two random variables x_1 and x_2 with a joint distribution $p(x_1, x_2)$ to help one better grasp Gibbs sampling. Gibbs sampling lets each variable's value change in turn. First we sample from $p(x_1|x_2)$ to get $x_1^{(t+1)}$; then, using the following formula, from $p(x_2 \left| x_1^{(t+1)}\right.)$ we get $x_2^{(t+1)}$:

$$x_1^{(t+1)} \sim p\left(x_1 \left| x_2^{(t)}\right.\right) \tag{3}$$

$$x_2^{(t+1)} \sim p\left(x_2 \left| x_1^{(t+1)}\right.\right) \tag{4}$$

Gibbs sampling can efficiently approximate the joint distribution $p(x_1, x_2)$ from which samples are finally taken by alternating updates of x_1 and x_2.

Usually in opinion analysis, we must identify possible topics from a lot of textual data. Assume that every document covers numerous subjects, and every subject consists of several words. By progressively assigning each document and topic, Gibbs sampling

can detect the possible themes of a document. Under this situation, let $w_{d,n}$ in document d be allocated to the topic $z_{d,n}$; then, Gibbs sampling updates the topic label $z_{d,n}$ for every word. The conditional probability has the formula as follows:

$$p\left(z_{d,n}=k\left|z_{-d,n}, w\right.\right) \propto \frac{\left(n_{d,k}^{-d,n}+\alpha\right)}{\left(N_d+K\alpha\right)} \cdot \frac{\left(n_{k,w_{d,n}}^{-d,n}+\beta\right)}{\left(n_k+V\beta\right)} \tag{5}$$

where N_d is the total number of words in topic k; $n_{d,k}$ is the number of words assigned to topic k in document d; $n_{k,w_{d,n}}$ is the frequency of word $w_{d,n}$ in topic k; α and β are hyper-parameters; V is the vocabulary size; k is the number of topics.

Apart from subject identification, Gibbs sampling finds application in sentiment analysis to assist with text sentiment trend identification (Gao et al., 2021). Gibbs sampling allows us, for instance, to deduce the sentiment label of every document. Given the joint distribution of the word $w_{d,n}$ and the sentiment label $e_{d,n}$, assuming that the sentiment label is $e_{d,n}$ we can update it. The conditional probability formula follows:

$$p\left(e_{d,n}=t\left|e_{-d,n}, w\right.\right) \propto \frac{\left(n_{d,t}^{-d,n}+\gamma\right)}{\left(N_d+T\gamma\right)} \cdot \frac{\left(n_{t,w_{d,n}}^{-d,n}+\delta\right)}{\left(n_t+W\delta\right)} \tag{6}$$

where N_d is the total number of words for the sentiment tag t; W is the size of the vocabulary; T is the number of sentiment tags; $n_{d,t}$ denotes the number of occurrences of the sentiment tag t in document d; $n_{d,t}$ denotes the number of occurrences of the word $w_{d,n}$ corresponding with the sentiment tag t, n_t is the total number of words for the sentiment tag t.

Gibbs sampling progressively updates the values of every variable therefore enabling the Markov chain to progressively converge to the target distribution (Green et al., 2015). Successive updates of the conditional probabilities drive this convergence; finally, the target distribution is the approximate one we require. Usually, multiple rounds are needed to guarantee the effectiveness of the method so that Gibbs sampling may converge stably.

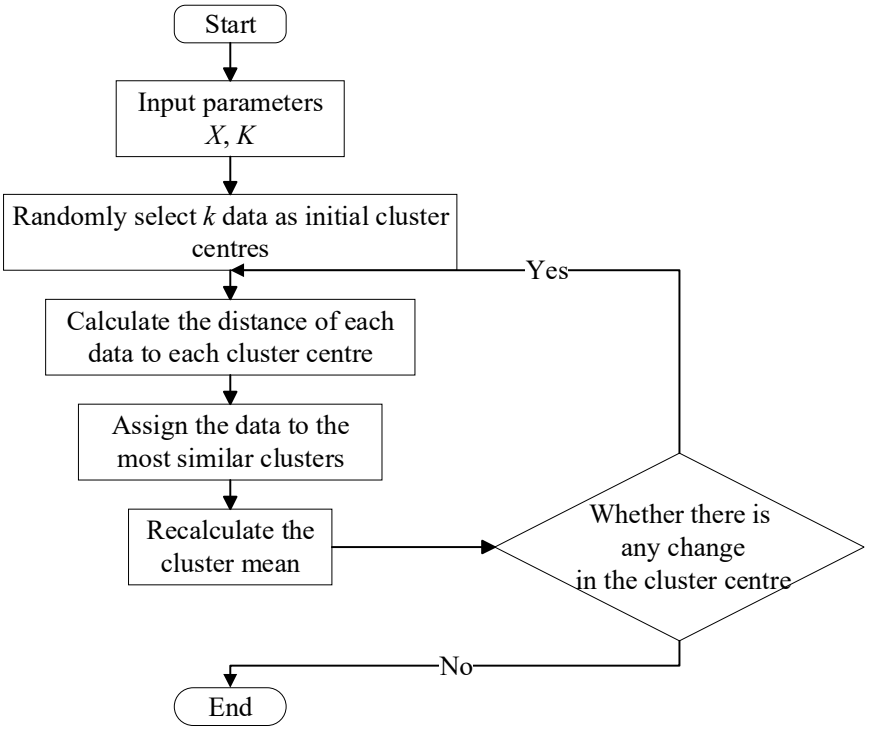
Finally, Gibbs sampling is a useful sampling technique having several uses in fields including sentiment analysis, topic modelling, and opinion analysis. Gibbs sampling is able to generate samples from intricate probability distributions by iteratively updating the conditional distribution of every variable, therefore exposing for us the fundamental patterns, themes, and sentiment elements in textual data.

2.2 Clustering algorithms

Clustering is a widely used data mining tool in the study of ideological opinion in the social new media environment that can efficiently extract possible themes, sentiments, or opinion tendencies from a big volume of text data (Xing et al., 2022). Clustering techniques assist to expose the fundamental structure in the data by grouping like data points. Among several clustering techniques, the K-means clustering algorithm is most often applied in the analysis of text data, particularly in the study of ideological opinion, which can so clearly identify the variations in opinion among several groups. Thus, this work selects K-means clustering technique and aggregates its benefits to investigate the properties of ideological public opinion in social new media.

Common unsupervised learning method extensively applied in the analysis of ideological beliefs in the social new media environment is K-means clustering (Zhang and Peng, 2024). See Figure 1 to understand its fundamental objective: partition the dataset into K clusters such that the similarity of samples within clusters is maximised and the variations between clusters are maximised.

Figure 1 K-means clustering algorithm



K-means is fundamentally based on dividing the data so that the centre of the cluster in which each data point is located is minimally far from every other data point. Assume we have N samples from a sample set $\{x_1, x_2, \ldots, x_N\}$ and every sample is a d-dimensional vector x_i that may be written as:

$$x_i = \left(x_{i1}, x_{i2}, \ldots, x_{id}\right) \tag{7}$$

The K-means clustering method aims to partition these data into K clusters whereby the mean of all the samples within each cluster can be stated. The centre μ_k of each cluster can therefore be expressed.

Through minimising the following objective function, the K-means algorithm generates clusters:

$$J = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \left\| x_i - \mu_k \right\|^2 \tag{8}$$

where ||.|| shows the Euclidean distance; N_k is the sample count in cluster k; μ_k is the centre of cluster k. The objective function's meaning is to decrease the total of the distances of the samples to the centres of the clusters to which they belong, therefore attaining clustering.

There three key parts to the K-means algorithm: initialising the cluster centres, assigning samples to the closest clusters, and updating the cluster centres (Ikotun et al., 2023). K samples are first chosen at random as initial cluster centres. Based on the position information of the current cluster centre, the method then computes the distance from each data point to all cluster centres in every iteration and assigns every sample to the cluster with the closest distance. We specifically determine the distance of every data point xi to the cluster centre μ_k and choose the cluster to which it most certainly belongs using the following formula:

$$k^* = \arg\min_{k'} \left\| x_i - \mu_{k'} \right\| \tag{9}$$

where $||x_i - \mu_{k'}||$ is the Euclidean distance between sample x_i and the cluster centre μ_k; k^* is the index of the cluster closest to sample x_i.

Then, given k^*, we can designate sample x_i to the matching cluster:

$$r_{ik} = \begin{cases} 1 & \text{if } k = k^* \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where r_{ik} is the indicator function showing whether data point belongs to cluster k or not.

The centre of every cluster is then updated; so, the mean value of all the cluster's samples defines the centre of the cluster:

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \tag{11}$$

where C_k is the collection of samples in cluster k; $|C_k|$ is the cluster's total count. This procedure keeps on until the maximum number of iterations is attained or until the cluster centre changes less than a specified threshold.

Minimising the objective function J – that is, the distance of every sample to the cluster to which it belongs – is the main goal of K-means clustering during algorithm iterations. The cluster centres and sample assignment is continuously changed to do this, thereby progressively bringing the local optimal solution. The mean values of the cluster's members are changed following every iteration to generate fresh cluster centres.

K-means clustering finds great use in data processing of social new media, where the text input is vectorised and the K-means algorithm analyses the text. Usually, initially employed to convert text input into vector representations, TF-IDF or Word2Vec is then the K-means algorithm used to cluster these vectors. Potential themes, feelings or opinion directions in the text data can be found by means of the clustering results, so offering important material for opinion analysis.

3 Modelling framework and integration

This work introduces a Gibbs sampling-based clustering recognition approach to precisely cluster and sentiment analyse the ideological public opinion in the social new

media environment. GibbsCluster is the model name; its basic concept is to identify the opinion data by clustering and additional mining of the sentiment trend in the text by combining the Gibbs sampling algorithm with conventional clustering algorithms, such K-means. Gibbs sampling is a fast a posteriori inference technique that uses random sample to optimise the cluster centre, hence improving the accuracy and resilience of cluster recognition.

Data are typically of great dimensions and complexity in the social new media environment, and noise and redundant information abound there as well. Thus, the architecture of the GibbsCluster model emphasises on making full use of information from many sources by means of cooperative work across several modules to provide accurate and efficient opinion analysis. The model's framework consists in a pre-processing module, a feature extraction module, a cluster identification module, a sentiment analysis module, and an evaluation and optimisation module. Through the combination of multimodal data, the modules cooperate to finish data cleaning, feature extraction, cluster recognition and sentiment analysis at several phases, and lastly increase the general performance of the model.

3.1 Pre-processing module

The pre-processing module aims to clean and normalise the raw data so that it is fit for later analysis as data in social new media sometimes consists of a lot of noise and pointless information. Eliminating deactivated words, punctuation marks and special characters, and using stemming extraction on the text are part of the pre-processing activities. We express the text data in this module using the TF-IDF (word frequency-inverse document frequency) technique. This approach allows us to convert the text input into numerical vectors for additional handling rather efficiently.

Every text t_i has a matching TF-IDF representation:

$$TF\text{-}IDF\left(t_{i,j}\right) = TF\left(t_{i,j}\right) \cdot \log\left(\frac{N}{DF\left(t_j\right)}\right) \tag{12}$$

where $t_{i,j}$ is the j^{th} word in the text; $TF(t_{i,j})$ is the frequency of the word in the text; $DF(t_j)$ is the number of documents having the word; N is the overall count of the documents.

For every text, TF-IDF allows us to create a high-dimensional vector ready for the next feature extraction and clustering study.

3.2 Feature extraction module

The feature extraction module's goal is to identify sentiment analysis and clusterable relevant features from the cleaned text data. We vectorise the textual data using Word2Vec technique. Every word can be expressed as a vector of fixed dimensions using Word2Vec, which also allows to capture the semantic interactions among words.

Assume the text comprises M words and that the word vector of every word w_j is $v(w_j)$. Summing all the word vectors in the text then yields the feature vector $v(t_i)$ of the entire text:

$$v(t_i) = \sum_{j=1}^{M} v(w_j) \tag{13}$$

where $v(t_i)$ is the vector representation of the text t_i and $v(w_j)$ is the Word2Vec vector of the word w_j. We thereby map every text to a low-dimensional vector space, which can more successfully capture the semantic content in the text.

3.3 Cluster recognition module

The core component of the GibbsCluster model is the cluster recognition module, which aims to cluster opinion data depending on textual aspects. We optimise the clustering process with Gibbs sampling method. By sampling from the posterior distribution, Gibbs sampling can update the cluster centres, hence enhancing the accuracy and stability of clustering.

Assuming a set of cluster centres in K-means clustering, μ can be stated as:

$$\mu = \{\mu_1, \mu_2, \ldots, \mu_K\} \tag{14}$$

We want Gibbs sampling to update the centre of every cluster. The posterior distribution of cluster centres μ_k can be stated at every sampling as:

$$p(\mu_k | X) \propto L(\mu_k | X) \cdot p(\mu_k) \tag{15}$$

where $p(\mu_k)$ is the prior distribution of the cluster centre; X is the feature set of all samples; $L(\mu_k|X)$ indicates the probability function of data point X given the cluster centre μ_k.

Gibbs sampling helps the clustering findings to converge to a locally optimal solution by repeatedly changing the cluster centres and sample assignments, therefore enhancing the efficacy of cluster analysis.

3.4 Sentiment analysis module

The sentiment analysis module aims to investigate the sentiment inclination of every cluster in order to assist in public opinion sentiment distribution revealing process. Using a sentiment analysis model, e.g., sentiment lexicon-based analysis or LSTM model – in this module we classify the sentiment of the text in every cluster depending on the text attributes of every cluster.

The sentiment analysis model computes the sentiment label $s(t_i)$ of the text by the following formula assuming $v(t_i)$ is the feature vector of text t_i:

$$s(t_i) = sign(W^T v(t_i) + b) \tag{16}$$

where W and b are the sentiment analysis model's parameters and sign() is a sign function denoting either positive, negative, or neutral sentiment classification of the text.

This module allows us to give sentiment labels to every cluster's words, therefore enabling a more thorough investigation of ideological opinions in social new media.

3.5 Multimodal data fusion

The GibbsCluster model integrates multimodal data fusion to combine information from several data sources, therefore improving the effect of clustering analysis. Apart from textual data, other aspects such user behaviour data, comment time, geographic information, etc. can also be incorporated. We improve the robustness of clustering analysis by weighted fusion of several kinds of data attributes.

The integrated feature $f(t_i)$ can be stated assuming the text feature is $v(t_i)$ as $u(t_i)$, user behaviour feature:

$$f(t_i) = \alpha \cdot v(t_i) + \beta \cdot u(t_i) \tag{17}$$

where α and β are weight coefficients expressing the relevance of user behaviour and text feature importance. At last, sentiment analysis and clustering will be fed from the combined features $f(t_i)$.

By means of multimodal data fusion, the GibbsCluster model can fully exploit several kinds of data to enhance the accuracy of sentiment analysis and clustering accuracy.

3.6 Evaluation and optimisation module

We choose three evaluation metrics: Silhouette coefficient, accuracy, and F1-score, which can reflect the clustering quality and sentiment analysis impact of the GibbsCluster model from various points of view, in order to evaluate the effectiveness of the model in the task of ideological opinion clustering totally.

First the quality of clustering is evaluated using Silhouette coefficient. Measuring the closeness of the data points to other points in their clusters and the dispersion from the closest clusters helps one to assess the clustering influence. More precisely, the contour coefficient $s(x_i)$ for every data point x_i is computed as:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\left(a(x_i), b(x_i)\right)} \tag{18}$$

where $b(x_i)$ is the average distance between point x_i and the closest cluster, $a(x_i)$ is the average distance between point x_i and other points in its cluster. Contour coefficient ranges span [–1, 1]. The closer the value is to 1, the stronger the clustering effect is; conversely, the clustering impact is less the closer the value is to –1.

Second, the sentiment analysis module's categorisation impact is graduated using the accuracy rate. The proportion of accurately predicted sentiment labels to the overall projected labels by the model defines the accuracy rate. Its computation formula is:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} \delta\left(s(t_i), \hat{s}(t_i)\right) \tag{19}$$

where $\delta(s(t_i), \hat{s}(t_i))$ is a function of indication with a value of 0 otherwise and 1 when the expected sentiment labels match the actual labels. In the sentiment classification job, accuracy indicates the general model correctness.

In sentiment analysis, F1-score is a widely used assessment tool that blends recall and accuracy. Particularly in cases of unbalanced samples, it gauges the general performance of the model in categorisation. F1-score has the formula:

$$F1\text{-}score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{20}$$

Among these, precision and recall assess respectively how many of all the truly positive examples are correctly detected and how many of the samples predicted by the model are correct. The model performs better the greater the value of F1-score.

We analyse the performance of the GibbsCluster model in cluster recognition and sentiment analysis holistically and objectively using these three evaluation indices; subsequently, we confirm its application effect in ideological opinion analysis.

By combining Gibbs sampling, K-means clustering, sentiment analysis and multimodal data fusion, the GibbsCluster model can essentially identify clusters and analyse sentiment patterns of ideological public opinion in the social new media environment. By means of the cooperative efforts of every module, the model achieves correct processing of high-dimensional public opinion data and produces outstanding clustering accuracy and sentiment recognition, so supporting public opinion analysis.

4 Experimental results and analyses

4.1 Datasets

This study mostly uses public opinion data on social new media platforms including Weibo, WeChat, Zhihu and other social platforms to create the experimental dataset. The choice of the dataset is to confirm the performance of sentiment analysis and clustering impact of the GibbsCluster model in the real social media environment. The collection comprises especially in two sections: sentiment label data and opinion text data.

From publicly available social media databases, we gathered a lot of opinion pieces about social hotspots, politics, education, economy, etc. These user-published text contents comprise comments, postings, debates, etc. In keeping with popular opinion in a real-world social media context, the text dataset consists of N = 50,000 entries encompassing a wide spectrum of areas and themes with great diversity and complexity.

Every opinion text is manually classified with sentiment labels, generally comprising three categories of positive, negative and neutral attitudes. We requested three seasoned annotators to annotate the data utilising several rounds of annotations to guarantee the correctness of sentiment labelling and consistent verification helped to settle the annotations' conflicts. Following consistency validation, the labelling results were eventually ascertained to guarantee high degree of trust in sentiment labels of the dataset.

The text was initially de-noised – that is, HTML tags, URLs, etc. – then it underwent lexical segmentation with deactivated words and irrelevant letters eliminated in the data preparation procedure. Following that, for next model training and evaluation, a vectorised form of every text was generated using the TF-IDF. Following pre-processing, 50,000 cleaned opinion texts overall were gathered from all text data.

We also split the dataset in a training set and a testing set for model development and testing. Ten thousand texts make up the exam set; 40,000 texts make up the training set. Data distribution ranges from 80% to 20% to provide both model evaluation validity and representativeness during the training phase.

Table 1 is a detailed description of the dataset.

Table 1 Dataset statistical information

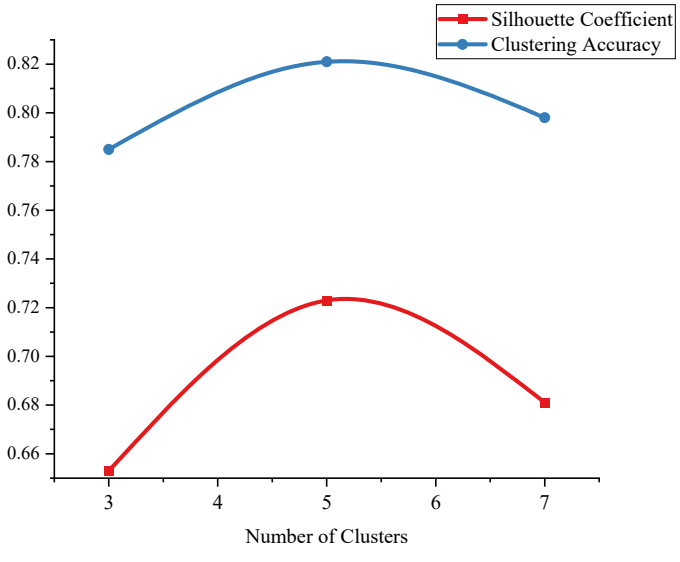
Dataset category	*Quantity*	*Description*
Public opinion text data	50,000 entries	Text data collected from platforms like Weibo, WeChat, and Zhihu, covering various social topics, politics, education, etc.
Sentiment label data	50,000 entries	Each opinion text is labelled with sentiment tags: positive, negative, and neutral.
Training set	40,000 entries	The training set comprises 80% of the total data, used for model training.
Test set	10,000 entries	The test set comprises 20% of the total data, used for model evaluation.

Apart from guaranteeing the variety and complexity of the data, the choice of the experimental dataset realistically reflects the traits of various subjects, emotions, and user behaviours in the social new media environment, so laying a strong basis for the validation of the GibbsCluster model.

4.2 Clustering effect experiments

We present a thorough evaluation of the GibbsCluster model's performance on social new media thinking opinion data in the clustering efficacy experiment. The major goals of the experiment are to evaluate the clustering quality with the Silhouette coefficient as the key assessment criterion and to validate the performance of the model under several number of clusters.

Figure 2 Clustering performance with different numbers of clusters (see online version for colours)



We set three distinct clustering numbers k = 3, 5, 7 matching varying numbers of opinion issues to run the studies. First, we found clusters for the data in the test set after training

the model with training set data. We computed the profile coefficient, clustering accuracy, and other model metrics in every experimental scenario to assess the impact of varying cluster counts.

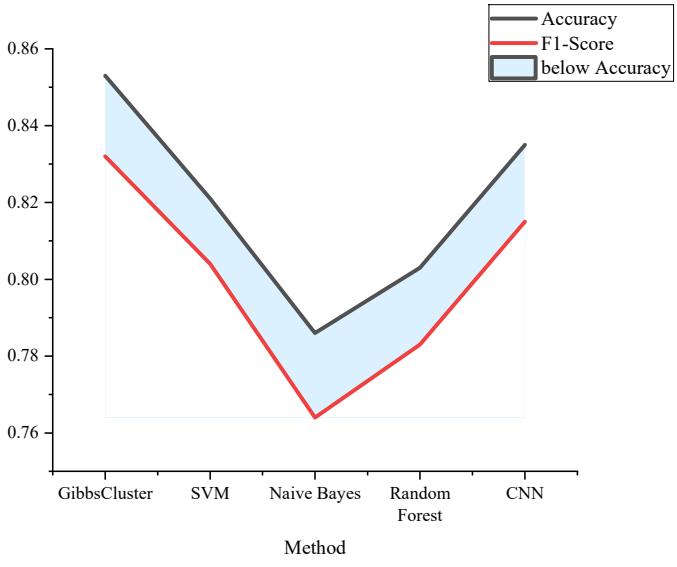
Figure 2 displays the outcomes of the studies on the clustering effect.

According to the experimental data, the contour coefficient and clustering accuracy find their ideal values at $k = 5$ when the number of clusters increases. The profile coefficient is specifically 0.723, which denotes that the clustering accuracy reaches 82.1% and the similarity inside the clusters is strong and the difference between various clusters is large. Whereas at $k = 7$ the clustering impact declines somewhat, the profile coefficient is 0.681, and the clustering accuracy is 79.8%; at $k = 3$ the clustering effect is really poor.

4.3 Sentiment analysis experiments

Our major objective in the sentiment analysis studies is to assess on thought-opinion data the efficiency of the GibbsCluster model for sentiment categorisation. First, the experiment assigns opinion pieces to several groups depending on clustering results. After that, we investigate if the sentiment inclination of the opinion texts in every cluster is positive, negative, or neutral by sentiment analysis. Accuracy and F1-score are the key evaluation measures we apply to assess sentiment analysis performance.

Figure 3 Performance comparison of different sentiment classification methods (see online version for colours)



Five sentiment classification techniques – the sentiment analysis module included with the GibbsCluster model, SVM, Naive Bayes, random forest, and CNN – were compared in our studies. Every technique guarantees fairness using the same training and test data. Particularly in the accuracy and F1-score of sentiment classification, which are superior than other conventional approaches, the experimental findings reveal that the GibbsCluster model performs effectively in sentiment analysis.

Figure 3 presents the outcomes of the sentiment analysis studies.

With an accuracy of 85.3% and an F1-score of 0.832, the GibbsCluster model clearly performs well from the experimental data in both accuracy and F1-score. This shows that GibbsCluster not only precisely evaluates the sentiment categories in the sentiment analysis task but also efficiently captures the actual sentiment distribution, so doing significantly better than other models.

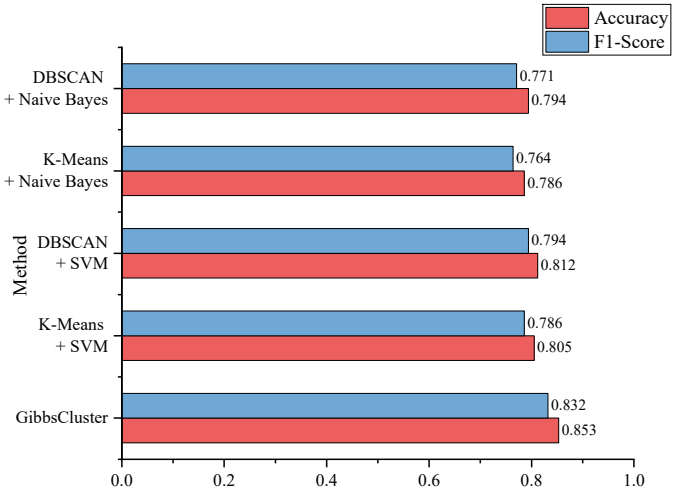
Other conventional sentiment classification techniques, such SVM, Naive Bayes, and random forest, on the other hand, show somewhat less performance in all metrics. For instance, Naive Bayes has an accuracy of 78.6% and an F1-score of 0.764 whereas SVM boasts an accuracy of 82.1% and an F1-score of 0.804. Although CNN falls short of GibbsCluster in terms of F1-score (0.815) and Accuracy (83.5%), it is rather near to both.

In sentiment analysis, together with the experimental data, we find that the GibbsCluster model performs noticeably better than conventional sentiment classification systems. The results support later multi-task tests strongly and confirm the effectiveness of the GibbsCluster model in opinion analysis.

4.4 Model comparison experiments

We intend to test the general performance of the GibbsCluster model against other common clustering and sentiment analysis methods in the processing of thought-opinion data in our model comparison experiments. We chose four well-known clustering and sentiment analysis methods for this aim for comparison: the GibbsCluster model, the DBSCAN clustering method, and the conventional K-means clustering approach. To guarantee fairness, every model was developed and tested using the same dataset. We assess the models' performance using accuracy and F1-score.

Figure 4 Comparison of model performance in clustering and sentiment analysis (see online version for colours)



In the experiments, sentiment analysis came first then K-means and DBSCAN techniques were applied for the clustering phase. Conversely, the GibbsCluster model optimises overall by combining sentiment analysis with clustering chores. Especially in terms of accuracy and F1-score performance, the experimental findings reveal that the

GibbsCluster model beats both the conventional clustering method and the independent sentiment analysis model.

Figure 4 present the outcomes of the model comparison studies.

With regard to accuracy (85.3%) and F1-score (0.832), the GibbsCluster model clearly beats the others based on trial results. On accuracy and F1-score, K-means and DBSCAN clustering techniques combined with a sentiment analysis model (SVM or Naive Bayes) did poorly. With an Accuracy of 78.6% and an F1-score of 0.764, the K-means + sentiment (Naive Bayes) technique especially performs most poorly.

Though in some cases the DBSCAN approach can identify clusters with higher density and the accuracy (81.2%) and F1-score (0.794) after merging with the sentiment analysis are somewhat improved compared to K-means, it still falls short to the level of the GibbsCluster model.

First, we verified the performance of the GibbsCluster model in the clustering task by means of the clustering effectiveness experiment; then, by means of the sentiment analysis experiment, we assessed the model's performance in the sentiment classification task; last, the model comparison experiment combined clustering and sentiment analysis to holistically evaluate the performance of the GibbsCluster model in the comprehensive task. By means of these successive experimental designs, we not only confirm the benefits of the GibbsCluster model in particular tasks but also show its universal effectiveness in practical settings.

5 Conclusions

The GibbsCluster model is proposed in this work to address sentiment analysis of ideological public opinion data in the social new media environment as well as clustering issues. GibbsCluster offers a creative way to monitor and evaluate ideological public opinion by efficiently automating the classification and sentiment analysis of public opinion data by merging the Gibbs sampling method with the K-means clustering algorithm.

Though it has several restrictions, the GibbsCluster model has shown good performance in many studies. For applications on large-scale datasets in particular, the model might have some computational overhead issues in face of very huge data quantities.

We intend to investigate the following two points of interest more thoroughly in next projects:

1 Optimising computational efficiency: Given large-scale datasets, the GibbsCluster model may now suffer severe computing overheads. By means of optimal algorithm implementation and parallel computing or distributed processing techniques for better applicability in large data environments, the computational efficiency of the model can be raised in the future.

2 Online learning and real-time analysis: Opinion data are dynamically changed in the social new media environment, so sentiment analysis and opinion monitoring depend critically on real-time. Future studies can investigate the online learning capabilities of the GibbsCluster model so that it can dynamically modify the clustering findings and sentiment categorisation during the data flow and adapt to the evolving opinion data.

Acknowledgements

This work is supported by the Research Project of Humanities and Social Sciences of the Ministry of Education of China (No. 23YJAZH168), the Humanities and Social Sciences Research Projects in Colleges and Universities in Jiangxi Province (No. SZZX23176), and the Special Funding for Basic Research Expenses for Central Government Department-Affiliated Universities (No. SKYZ2024036).

Declarations

All authors declare that they have no conflicts of interest.

References

Bansal, M., Goyal, A. and Choudhary, A. (2022) 'A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning', *Decision Analytics Journal*, Vol. 3, p.100071.

Dobbelaere, M.R., Plehiers, P.P., Van de Vijver, R. et al. (2021) 'Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats', *Engineering*, Vol. 7, No. 9, pp.1201–1211.

Gao, C., Zeng, J., Wen, Z. et al. (2021) 'Emerging app issue identification via online joint sentiment-topic tracing', *IEEE Transactions on Software Engineering*, Vol. 48, No. 8, pp.3025–3043.

Gnanasekaran, N. and Balaji, C. (2013) 'Markov chain Monte Carlo (MCMC) approach for the determination of thermal diffusivity using transient fin heat transfer experiments', *International Journal of Thermal Sciences*, Vol. 63, pp.46–54.

Green, P.J., Łatuszyński, K., Pereyra, M. et al. (2015) 'Bayesian computation: a summary of the current state, and samples backwards and forwards', *Statistics and Computing*, Vol. 25, pp.835–862.

Ikotun, A.M., Ezugwu, A.E., Abualigah, L. et al. (2023) 'K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data', *Information Sciences*, Vol. 622, pp.178–210.

Jelodar, H., Wang, Y., Yuan, C. et al. (2019) 'Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey', *Multimedia Tools and Applications*, Vol. 78, pp.15169–15211.

Kapoor, K.K., Tamilmani, K., Rana, N.P. et al. (2018) 'Advances in social media research: past, present and future', *Information Systems Frontiers*, Vol. 20, pp.531–558.

Linders, D. (2012) 'From e-government to we-government: defining a typology for citizen coproduction in the age of social media', *Government Information Quarterly*, Vol. 29, No. 4, pp.446–454.

Loader, B.D. and Mercea, D. (2011) 'Networking democracy? Social media innovations and participatory politics', *Information, Communication & Society*, Vol. 14, No. 6, pp.757–769.

Mostafa, M.M. (2013) 'More than words: social networks' text mining for consumer brand sentiments', *Expert Systems with Applications*, Vol. 40, No. 10, pp.4241–4251.

Sharma, C., Sharma, S. and Sakshi (2022) 'Latent Dirichlet allocation (LDA) based information modelling on blockchain technology: a review of trends and research patterns used in integration', *Multimedia Tools and Applications*, Vol. 81, No. 25, pp.36805–36831.

Traver, V.J. (2010) 'On compiler error messages: what they say and what they mean', *Advances in Human-Computer Interaction*, Vol. 2010, No. 1, p.602570.

Xing, Y., Wang, X., Qiu, C. et al. (2022) 'Research on opinion polarization by big data analytics capabilities in online social networks', *Technology in Society*, Vol. 68, p.101902.

Xu, S., Liu, J., Chen, K. et al. (2022) '[Retracted] research on the communication path of public opinion in university ideological and political network for big data analysis', *Journal of Sensors*, Vol. 2022, No. 1, p.8354909.

Zhang, H. and Peng, Y. (2024) 'Image clustering: an unsupervised approach to categorize visual data in social science research', *Sociological Methods & Research*, Vol. 53, No. 3, pp.1534–1587.