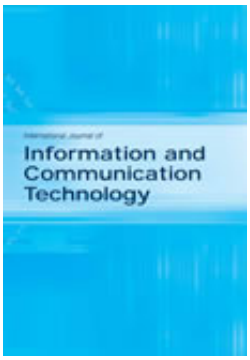
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

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

The value orientation clustering analysis based on topic models in the social network environment

Huimin Wang

Article History:

Received:	10 February 2025
Last revised:	19 February 2025
Accepted:	19 February 2025
Published online:	16 April 2025

The value orientation clustering analysis based on topic models in the social network environment

Huimin Wang

School of Marxism,
Gansu Police College,
Lanzhou 730299, China
Email: wymm1018@163.com

Abstract: The amount of user-generated data is growing with the fast expansion of social networks and e-commerce platforms; so, how to recognise and evaluate user values from these enormous amounts becomes a crucial study issue. In this study, a hybrid technique framework based on the combination of LDA topic model and DBSCAN clustering algorithm is provided for effective analysis of user values. First, the LDA model mines the possible themes of user-generated text data; then, the DBSCAN clustering method is applied to classify the behavioural traits of several user groups. Strong scalability and universality as well as accurate identification and classification of user values are shown by experimental validation on two real datasets of the proposed method. Better results in multi-dimensional user values analysis are obtained by the hybrid technique based on topic model and cluster analysis than by the conventional single model approach.

Keywords: social networks; user values; topic modelling; latent dirichlet allocation; LDA; DBSCAN; cluster analysis.

Reference to this paper should be made as follows: Wang, H. (2025) 'The value orientation clustering analysis based on topic models in the social network environment', *Int. J. Information and Communication Technology*, Vol. 26, No. 8, pp.19–34.

Biographical notes: Huimin Wang obtained her Bachelor's degree from the Northwest Normal University in 2008. She is currently an Associate Professor in the School of Marxism at Gansu Police College. Her research interests include online ideological and political education, public security characteristic culture, and public opinion prevention and control.

1 Introduction

With the rapid expansion of the internet, social networks and e-commerce platforms have become essential venues for social and corporate activities globally (Mata and Quesada, 2014). These platforms generate vast volumes of data every day, including users' social interactions, content uploading, consumption habits and comment feedback. Social networks give a virtual interactive place for users to express their thoughts, emotions and values by posting statuses, comments, likes and sharing material. Conversely, e-commerce sites show consumers' tastes and values by means of their purchasing behaviour, product reviews, and browsing patterns (Benlian et al., 2012). Not only does

knowing user values underlying these data help tailored recommendation systems, sentiment analysis and ad targeting, but it also gives insights for platform administrators to enable them to maximise their goods and services.

Based on their cultural background, social experiences, and personal feelings, users values are the fundamental attitudes and opinions of people on social, political, and economic spheres. Users' values are conveyed on social networks and e-commerce sites by means of several interactive activities including text, photographs, and videos. But given the diversity and complexity of the material users post on social media, precisely spotting and deciphering their values has grown difficult. Users' values are dynamic, changing with time and shaped by their surroundings, social events, and personal development.

Usually depending on manually tagged data and rule-driven models, traditional techniques of user value analysis suffer low efficiency and poor accuracy when dealing with large-scale and dynamic social network data (Abkenar et al., 2021). Methods based on automated analysis have progressively taken front stage as deep learning and natural language processing technology evolve. Many studies have tried to expose users' emotional tendencies and value profiles on social media by applying methods such sentiment analysis, subject modelling and cluster analysis (Yue et al., 2019).

Sentiment analysis, a popular method in current research, has been extensively applied to examine user behaviour on social networks and e-commerce platforms (Huang et al., 2023). Sentiment analysis often classifies the emotions, (e.g., positive, negative, neutral) of the text data uploaded by users, therefore identifying their emotional tendencies. Among common sentiment analysis techniques include deep learning, machine learning, and dictionary-based approaches (Yang et al., 2020). Deep learning-based sentiment analysis techniques have great benefits in handling complex emotional expressions, according to studies; but, their strong reliance on data and the possibility that they might not be able to sufficiently reflect the depth and variety of users' emotions in some situations define their drawbacks as well.

Furthermore, as an unsupervised learning approach, topic models have produced amazing outcomes in possible topic detection in text data. By modelling the possible distribution of words in the documents, latent dirichlet allocation (LDA), the most traditional topic model, may efficiently extract possible themes in documents and offer important data for user behaviour research (Jelodar et al., 2019). LDA finds great application in opinion monitoring, social media analysis, and news clustering. Many researches have mined the primary themes or areas of interest individuals bring by analysing user-generated information grounded on LDA models. Though the LDA model is great in topic discovery, it mostly depends on the surface information of the text and lacks the ability to capture the deeper meaning of the text and the user's emotions, so unable to totally meet the exact analysis of user values (Egger and Yu, 2022).

Furthermore, extensively applied in the challenge of social network user classification are cluster analysis approaches. By means of behavioural data analysis, common clustering techniques such K-means and DBSCAN classify users into several groups (Monalisa and Kurnia, 2019). Density-based spatial clustering of applications with noise, or DBSCAN, is a density-based clustering method capable of efficiently managing noisy data and spotting user groups with comparable behaviour patterns. Nevertheless, the clustering method also has certain restrictions, particularly with high-dimensional data; hence, selecting the suitable clustering method and parameters becomes very difficult.

The complexity and dynamism of the data structure of social networks and e-commerce platforms mean that even if these approaches have already shown some results in user behaviour analysis of social networks and e-commerce platforms, they still present many difficulties in handling the variety of user values and the complexity of emotions. Therefore, a more accurate and efficient user values analysis depends on how to employ theme models and cluster analysis in combination with users' behavioural patterns and emotional tendencies. This is still an essential issue to be solved.

This paper intends to provide a hybrid method based on topic modelling and cluster analysis for analysis of users' values in social networks and e-commerce platforms. The major objective of this work is to accomplish automated classification and mining of user values on social networks and e-commerce platforms by merging the LDA topic model with the DBSCAN clustering method.

This work offers the following novelties:

1 A hybrid approach combining thematic modelling and cluster analysis. This research provides a methodological framework based on the combination of LDA topic model and DBSCAN clustering algorithm. Through cluster analysis, this hybrid approach may not only identify possible subjects from user-generated text data but also efficiently classify users' behavioural patterns and values. This method closes the distance that conventional sentiment analysis and topic modelling cannot clearly expose the variety of user values.

2 Experimental validation based on multiple data sources. Two genuine datasets (Amazon product review dataset and Twitter sentiment dataset) are used in this work for experimental validation and the efficacy of the suggested strategy on several data sources is shown. The general adaptability of the hybrid technique in useful applications is evaluated by comparative studies of several datasets.

3 Extensibility and flexibility of methodological framework. Strong scalability and adaptability allow the suggested hybrid architecture to mix topic models with clustering techniques. Combining more deep learning techniques, image recognition, audio data, and other multimodal information will enable more thorough user value analysis in further studies.

These developments not only enhance the theoretical research of user value analysis but also offer fresh concepts and technical solutions for the actual implementation of social networks and e-commerce platforms.

2 Relevant technologies

2.1 Analysis of social networks and values

Particularly among groups, information sharing in social networks often follows certain trends; so, the transmission of values is a complicated process (Christakis and Fowler, 2013). The flow of information and the way values are passed through the exchanges, social contacts, and emotional resonance of people in the network. Analysing values in social networks thus calls not only for considering the information flow between individuals but also for combining the features of group behaviour with network structure.

Assuming that a social network comprises of a set of nodes V, one may represent V as:

$$V = \{v_1, v_2, ..., v_n\} \tag{1}$$

Simplified models allow us to represent the neighbourhood matrix of a social network by A, where a_{ij} is the strength of the link between nodes v_i and v_j. The value spreading dynamics of each node can be simulated using the following equation if network members disseminate a specific value xi via interactions:

$$x_i(t+1) = \alpha \sum_{j \in N(i)} a_{ij} x_j(t) + \beta x_i(t) \tag{2}$$

where $x_i(t)$ denotes the value state of node v_i at moment t; $N(i)$ denotes the set of nodes adjacent to node v_i; α is the weight coefficient indicating the strength of information distribution among nodes; β is the self-feedback coefficient indicating the ability of the values of an individual node to self-maintain when there is no outside influence.

The unidirectional character of value propagation in social networks hides the crucial relevance of node-to-node influence and feedback (Kaligotla et al., 2022). Graphs theoretically let us compare groups depending on the degree of similarity between their nodes. One may define the resemblance $S(v_i, v_j)$ between two nodes v_i and v_j as follows:

$$S(v_i, v_j) = \frac{A_{ij}}{\sqrt{\sum_k A_{ik}^2 \sum_l A_{jl}^2}} \tag{3}$$

Usually, their connectivity patterns help one to determine their similar behaviour or values. This similarity metric offers a framework for later value clustering studies as well as a means to fairly capture the homogeneity of values inside a group.

Sometimes outside elements like political atmosphere and cultural background affect the way values are passed on via social media. These elements influence not only the development of personal values but also help to decide if some ideals will be generally embraced in a given group. Consequently, while simulating the spread of values in social networks, it is usually essential to include the impact of outside information sources. Assuming that the impact of outside elements on a person's values may be expressed by z_i, the value propagation of a node can be changed as:

$$x_i(t+1) = \alpha \sum_{j \in N(i)} a_{ij} x_j(t) + \beta x_i(t) + \gamma z_i \tag{4}$$

where z_i marks the external influence on node v_i and γ is the weight coefficient of external influence.

In social networks, the development of group behaviours is sometimes the outcome of several and complicated elements (Newman and Park, 2003). The group dynamics model helps us to better grasp the evolution process of personal and collective values in the network. Under this paradigm, the values of the group as a whole also affect the values of individuals in addition to the adjacent nodes. Assume that the following equation adequately explains the value evolution of node v_i:

$$x_i(t+1) = \sum_{j \in N(i)} w_{ij} x_j(t) + \delta \left(\frac{1}{|N(i)|} \sum_{j \in N(i)} x_j(t) + x_i(t) \right) \tag{5}$$

where w_{ij} is the weight of node v_i and node v_j, and δ is the group impact coefficient, which reflects the degree of influence of the overall values of the group on the values of individuals.

These models provide a mathematical framework for value transmission in social networks and help us better understand the behavioural characteristics of groups and individuals in social networks.

2.2 Thematic modelling and cluster analysis methods

Subject models (e.g., LDA) enable us in social network analysis to identify the underlying subject structure from a lot of user comments or posts. Every document (or post) in LDA combines several themes derived from a probability distribution of a set number of phrases.

Every paper generates as follows: first a topic is chosen from a topic distribution; then, a vocabulary from the vocabulary distribution of this topic is chosen (Hagen, 2018). The LDA model's generating process can be characterised with the following condensed formula:

$$p\left(w_{i,j} \left| z_{i,j}, \phi_{z_{i,j}} \right.\right) = \phi_{z_{i,j}, w_{i,j}} \tag{6}$$

where $\phi_{z_{i,j}, w_{i,j}}$ is the likelihood of the j^{th} word $w_{i,j}$ under the topic $z_{i,j}$; $w_{i,j}$ indicates the j^{th} word in the document d_i, $z_{i,j}$.

Maximising the likelihood of document creation by inferring the distribution of topics θ_i of a document and the distribution of words ϕ_k of each subject shows the possible themes covered in the social network, hence guiding the LDA model's central purpose. One can simplify LDA's optimisation goal as follows:

$$L = \max_{\theta, \phi} \prod_{i=1}^{M} p\left(d_i \left| \theta_i, \phi \right.\right) \tag{7}$$

Therefore, we wish to maximise the possibility of generating all documents to deduce the subject distribution of documents and the lexical distribution of themes, so understanding the interests and values of users in a social network.

By grouping like objects, cluster analysis – a unsupervised learning technique – helps us find several groupings in a social network. Based on their behaviour, topic debates, or other characteristics, cluster analysis can find groupings of consumers with like values. Among common clustering techniques are DBSCAN, hierarchical clustering, and K-means clustering (Popat and Emmanuel, 2014). Consider K-means clustering: imagine we have a user feature set X; then X can be written as:

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

Every feature vector x_i stands for a user's interest or behaviour. With the intention of reducing the sum of the distances from each data point to the cluster in which it is placed, the K-means clustering algorithm first chooses K clustering centres $\mu_1, \mu_2, \ldots, \mu_K$ and then assigns each user to the nearest cluster based on their distance from the clustering centre:

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \left\| x_i - \mu_k \right\|^2 \tag{9}$$

where μ_k is the clustering centre of the k^{th} cluster and Ck is the collection of users in the k^{th} cluster.

The process of K-means algorithm is as follows: first, K cluster centres are chosen at random; next, based on the distance of every user x_i to every cluster centre, the cluster C_k to which each user x_i belongs is ascertained; subsequently, the centre μ_k of every cluster is changed to be the mean of the characteristics of all the users inside that cluster:

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \tag{10}$$

At last, the foregoing actions are carried out until the centre of clustering remains constant.

Cluster analysis results can enable us to distinguish several user groups in a social network. Clustering users helps one to identify whether groups of them have comparable inclination to address a given issue or value. For instance, some users can be more focused on social justice problems while others would mostly talk on environmental ones (Kumar and Reddy, 2017). Clustering techniques help to expose these groups and guide next value studies.

Apart from K-means clustering, another often used clustering technique is DBSCAN, hierarchical clustering, Hierarchical clustering calculates the similarity between every pair of users and progressively merges or splits users depending on the similarity, therefore forming a tree structure. Hierarchical clustering can be shown by the following similarity matrix assuming Euclidean distance or cosine similarity measures of user feature similarity:

$$S_{ij} = 1 - \frac{\left\| x_i - x_j \right\|}{\max\left(\left\| x_i \right\|, \left\| x_j \right\| \right)} \tag{11}$$

where S_{ij} indicates the similarity between users i and j; $||x_i - x_j||$ is their Euclidean distance.

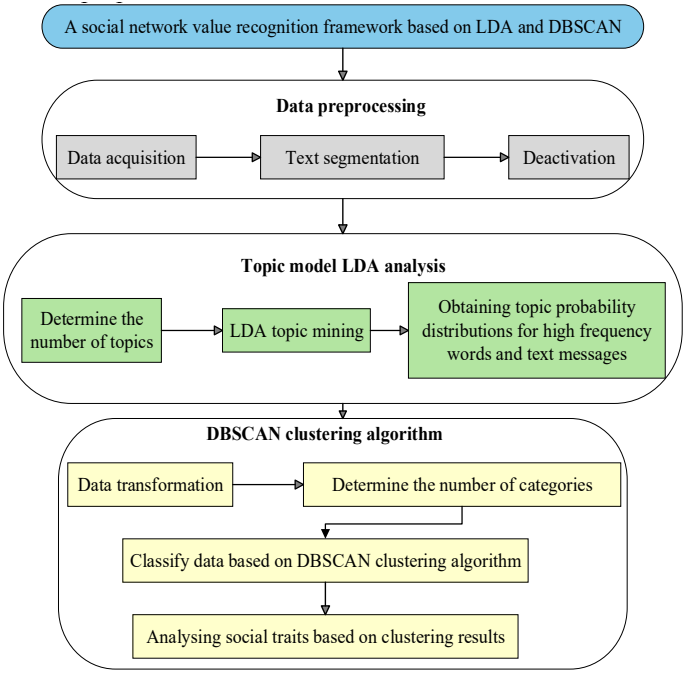
By defining the radius of the neighbourhood ϵ and the minimal number of samples, MinPts, DBSCAN clustering finds clusters using the concept of density rather than requiring an advance stated number of clusters (Mittal et al., 2019). DBSCAN uses the clustering criterion whereby a point is a core point and clustering will be extended from that point if the neighbourhood of a point has at least MinPts of points.

Cluster analysis's basic idea is to utilise similarity measurements to group consumers with like interests and values. Common similarity measures for social network data consist in content-based similarity, (e.g., cosine similarity) and behaviour-based similarity (e.g., Jaccard similarity). These similarity criteria enable us to cluster people in a social network thereby exposing the value variations among various groupings.

3 A framework for identifying social network values based on thematic modelling and cluster analysis

By means of combination theme modelling with cluster analysis, the framework seeks to mine and discover users' values from textual data in social networks, see Figure 1.

Figure 1 Structure of proposed model (see online version for colours)



There are four main basic elements to the framework overall:

1 Data pre-processing and feature extraction

Aiming at cleaning, denoising and characterising text data in social networks to give efficient inputs for later topic modelling and clustering analysis, data reprocessing is the fundamental step of the complete system. Operations including text cleaning, word splitting, deactivated word removal, and TF-IDF vectorising of text data by means of this formula define the pre-processing process:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t) \tag{12}$$

$$\text{TF}(t, d) = \frac{\text{Term frequency of } t \text{ in document } d}{\text{Total number of terms in document } d} \tag{13}$$

$$\text{IDF}(t) = \log\left(\frac{N}{\text{Number of documents containing term } t}\right) \tag{14}$$

where N is the overall count of the documents; t is a lexical item; d is a specific document.

This allows us to produce feature vectors for every document and offer useful data for next models.

2 Topic modelling (LDA)

Through statistical learning, LDA modelling is applied to mine the latent subjects in text. In LDA, it is hypothesised that every document is produced by several themes with varying probability distributions and that every subject is produced by distinct words with particular probability distribution.

One can see the LDA model's topic generating process by means of the following equation:

$$p(w \mid \theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} p\left(w_n \mid \phi_k\right) p(\theta \mid \alpha) \tag{15}$$

where α is the prior distribution of topics; w is the words in the document; θ is the topic distribution of the document; ϕ_k is the word distribution of the k^{th} subject; K is the number of topics.

Maximising this log-likelihood function allows the LDA model to estimate the topic distribution θ for every document and the word distribution ϕ_k for every topic.

3 Cluster analysis (DBSCAN)

Robust to noise, DBSCAN is a density-based clustering technique that detects the density of data points and so efficiently finds clusters of any kind.

DBSCAN is fundamentally defined by specifying core, border, and noise points. One can write its clustering guidelines by the following equation:

$$\text{CorePoint}(p) = \left|\{q \in D : \text{Dist}(p, q) \leq \epsilon\}\right| \geq \text{MinPts} \tag{16}$$

where ϵ is the radius threshold; MinPts is the lowest number of neighbourhood points; Dist(p, q) is the distance between point p and point q. Data points classified as core points have sufficient neighbours within a defined radius.

Regarding boundary points, the following criteria hold:

$$\text{BoundaryPoint}(p) = \left|\{q \in D : \text{Dist}(p, q) \leq \epsilon\}\right| \geq \text{MinPts} \tag{17}$$

This density connectivity shapes the DBSCAN clustering results and finally helps to define the limits and core components of the various clusters.

4 Fusion of themes and clustering results

Using LDA and DBSCAN in the framework allows one to define a subject feature for every cluster. We must somehow mix the thematic information gathered by LDA with the DBSCAN clustering results to execute the fusion.

Usually, one gets the clustered topic vectors by computing the topic distribution for every cluster using a weighted average technique.

One can represent the weighted average of the clustered topic features by means of the following equation:

$$z_{C_k} = \frac{1}{|C_k|} \sum_{d \in C_k} \theta_d \tag{18}$$

where z_{C_k} is cluster C_k's topic vector. $|C_k|$ is the quantity of papers in cluster C_k; θ_d is the subject distribution of document d.

In this sense, we can create a topic feature vector for every cluster, therefore aggregating the outcome of the topic model with the clustering results.

5 Visualisation and analysis of results

Techniques of visualisation can enable us to grasp the outcomes of topic distribution and clustering. Principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE) are common dimensionality reduction techniques able to map high-dimensional data into a two-dimensional space for simple analysis and display.

PCA dimensionality lowering formula:

$$X_{\text{PCA}} = XW \tag{19}$$

where W is the projection matrix reflecting the main component directions derived via eigenvalue decomposition and X is the original data matrix. These main component directions map the dimensionality lowered data X_{PCA} on.

Using t-SNE helps us to further lower the data dimensionality and preserve the local structural information for the high-dimensional data visualisation:

$$Y = \arg\min_{Y} \sum_{i \neq j} \left(p_{ij} - q_{ij} \right)^2 \tag{20}$$

where p_{ij} is the conditional probability distribution between points i and j in the high-dimensional space; q_{ij} is the conditional probability between points i and j in the low-dimensional space following dimensionality reduction.

6 Evaluation and optimisation

- Silhouette coefficient:

  One of the indicators of the quality of clustering is the silhouette coefficient; a value nearer 1 denotes better clustering and a negative value denotes poor clustering. Silhouette coefficient computes the similarity $a(i)$ between each data point and other points in the cluster and the similarity $b(i)$ between the data point and the closest cluster, therefore evaluating the closeness and separation of clusters.

  The formula is shown below:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{21}$$

  The average of all the data points determines the profile factor generally:

$$S_{\text{avg}} = \frac{1}{N}\sum_{i}^{N} S(i) \tag{22}$$

The average distance of data point i from other points inside the same cluster is $a(i)$ the average distance of the point from the closest other cluster is $b(i)$.

- Adjusted rand index (ARI):

  The agreement between the genuine labels and the clustering results is calculated with ARI. Its value falls between [–1, 1] corresponds to the better clustering outcome the closer to 1.

  ARI uses a formula like this:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \tag{23}$$

  where RI is the rand index; $E[RI]$ is the intended rand index; $\max(RI)$ is the rand index's highest value.

  The rand index RI can be computed from the following formula:

$$RI = \frac{TP + TN}{\binom{N}{2}} \tag{24}$$

  where TP is the total number of correctly matched pairs of samples in the cluster, TN is the total number of erroneously matched pairs of samples, and N is the total number of data points.

- Normalised mutual information (NMI):

  Larger numbers indicate better clustering; NMI gauges the degree of information shared between the true labels and the clusterering outcomes. NMI can be computed with this formula:

$$NMI(U, V) = \frac{I(U, V)}{\sqrt{H(U)H(V)}} \tag{25}$$

  where $H(U)$ and $H(V)$ are respectively the entropy computed for the clustering result and the true label correspondingly; $I(U, V)$ is the mutual information between them.

  Calculated as mutual information $I(U, V)$ is:

$$I(U, V) = \sum_{u \in U}\sum_{v \in V} p(u, v)\log\frac{p(u, v)}{p(u)p(v)} \tag{26}$$

  where $p(u, v)$ is the joint probability distribution of the clustering results and the true labels; $p(u)$ and $p(v)$ are respectively the marginal probability of the clustering results and the labels.

These three metrics can fairly represent the performance of the model in the clustering task and offer a whole evaluation of the model in terms of tightness, accuracy and information sharing of clustering, respectively.

4 Experimental results and analyses

4.1 Experimental data

Two datasets – the Twitter sentiment dataset and the Amazon product review dataset – are chosen for tests in this work in order to validate the efficiency of the suggested values clustering method based on topic modelling and cluster analysis. Covering several social platforms and online buying situations, these two datasets enable extensive textual information for this study, therefore enabling analysis of consumers' emotions and values from several angles.

Table 1 shows the dataset's overall fundamental information:

Table 1 Dataset statistical information

Dataset name	*Number of records*	*Feature dimensions*	*Label types*	*Data type*
Twitter sentiment dataset	100,000	10,000+	Three sentiment labels (positive, negative, neutral)	Text data with sentiment labels
Amazon product review dataset	500,000	8,000+	5 star rating (1 to 5 stars)	Text data with review ratings

In the Twitter dataset, sentiment labels; in the Amazon dataset, rating labels were normalised. Sentiment labels (positive, negative, and neutral) were changed to numerical labels (1, 0, –1) for the Twitter sentiment dataset; rating labels (1 to 5 stars) were changed to numerical labels for the Amazon product review dataset. Using the TF-IDF (word frequency-inverse document frequency) technique – which presents the text as sparse vectors – features are obtained.

Tasks involving the Twitter sentiment dataset and the Amazon product review dataset will be carried out: clustering the tweets or reviews using clustering algorithms, (e.g., K-means or DBSCAN), investigating the various groups of sentiment and value characteristics, and perform cluster analysis of values; extracting potential themes from the text by means of the LDA model and analyse the relationship between each theme and the sentiment labels or the rating labels. Rich empirical data can be supplied for the clustering study of user values in social networks and e-commerce platforms by means of examination of two datasets.

4.2 Experiments in analysing emotions and values based on thematic models

Using the Twitter sentiment dataset and the Amazon product review dataset, the experiment assesses the capacity of the model to extract sentiment and values in a social network environment. The LDA model analyses the abundance of textual data in every dataset to extract underlying themes and mix them with sentiment tags (Twitter) or rating tags (Amazon). The experiment aims to investigate the link between sentiment and themes as well as to assess the model's performance in clusterering analysis-based user value classification.

Figures 2 and 3 respectively demonstrate the findings of the tests on the two datasets.

Figure 2 Clustering evaluation metrics for DBSCAN on Twitter sentiment dataset (see online version for colours)

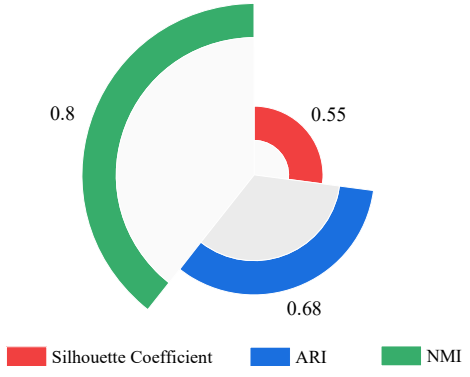
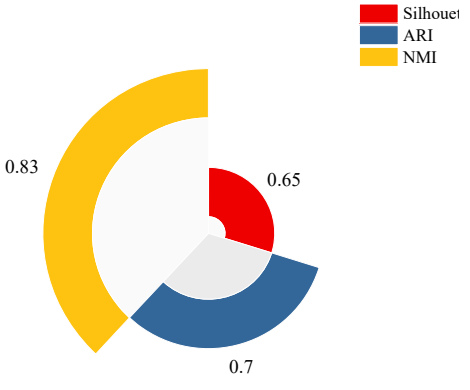


Figure 3 Clustering evaluation metrics for DBSCAN on Amazon product review dataset (see online version for colours)



The aforementioned results indicate that DBSCAN clustering can efficiently detect the distribution of themes between several sentiment labels or rating labels and shows better resilience on the two datasets. DBSCAN shows improved clustering results in the Twitter sentiment dataset especially by having higher profile coefficients and ARI values. Furthermore, DBSCAN performs especially well on NMI in the Amazon product review dataset, therefore demonstrating strong correlation between sentiment labels and derived topics. These results show that the sentiment and value analysis method based on LDA topic model paired with DBSCAN clustering has good potential for usage since it can efficiently extract users's sentiment information and underlying values from social network and e-commerce platform data.

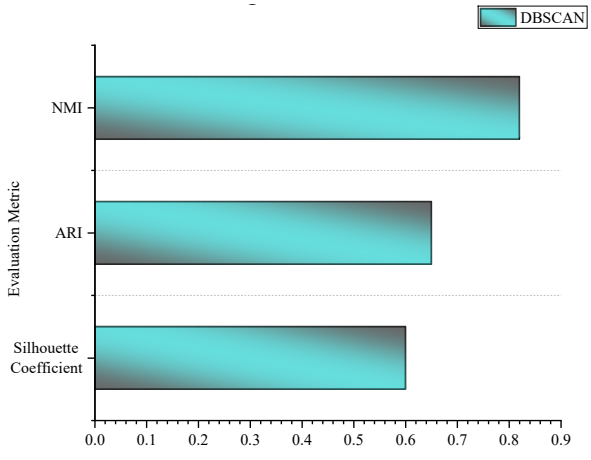
4.3 Experiments on classifying user values in social networks and e-commerce platforms based on cluster analysis

This experiment intends to utilise cluster analysis to classify, via social network and e-commerce platform user values. We investigated user underlying values by means of clustering techniques using the Twitter sentiment dataset and the Amazon product review

dataset. In this experiment, we automatically cluster users' interests and values using textual material instead of conventional labels for classification.

First standardised pre-processing – text cleaning, deactivation removal, word shape reduction, TF-IDF feature extraction – was applied to the datasets in the trials. Every dataset was then grouped with DBSCAN clustering method. Unlike pre-defined category labels, the clustering method finds possible user groups based on text similarities on its own initiative. At last, we examined the match between the clustering results and the real labels using measures like silhouette coefficient, ARI, and NMI in order to assess the clustering impact. Figures 4 and 5 respectively exhibit the experimental outcomes.

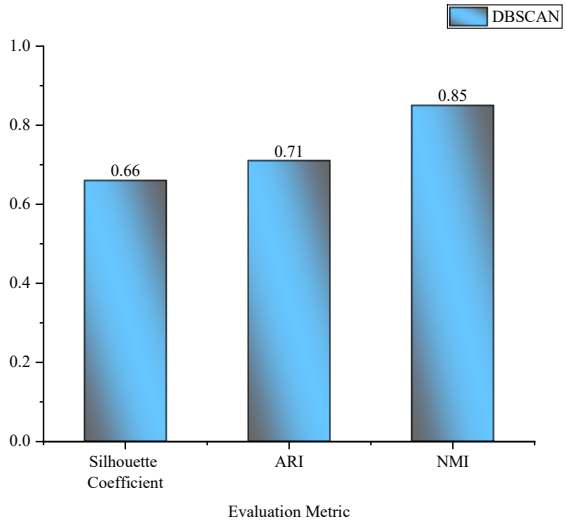
Figure 4 Evaluation metrics for DBSCAN clustering on Twitter sentiment dataset (see online version for colours)



DBSCAN clustering produces rather reasonable outcomes in the studies on the Twitter sentiment dataset. With a contour coefficient of 0.60, the clustering effect is medium, suggesting some degree of separation among the several emotional groupings. With an ARI of 0.65, DBSCAN can efficiently identify user groups with varying emotional inclinations and the clustering result fits the actual emotional labels rather nicely. With a N MI of 0.82, the clustering method is clearly able to more precisely capture the emotional information since the real labels show a high correlation with its outcome.

Experiments on the Amazon product review dataset show DBSCAN clustering's higher importance. Slightly higher than the Twitter dataset, the contour coefficient is 0.66, suggesting that the review contents and rating labels help to better classify the user groups and separate and coherent clustering findings. With an ARI of 0.71, the clustering findings show a great degree of fit with the real rating labels and can fairly represent the grouping of the reviews across the several rating intervals. With a N MI of 0.85, the clustering results and the degree of correlation between them show is really high. DBSCAN is more successful in classifying user values on this dataset since the NMI is 0.85, which indicates a strong correlation between the clustering results and the rating labels.

Figure 5 Clustering evaluation metrics for DBSCAN on Amazon product review dataset (see online version for colours)



Particularly on the NMI and ARI measurements, which show high values, indicating a great degree of match between the clustering results and the actual labels, the experimental results show that the DBSCAN clustering technique shows a better clustering effect on both the Twitter sentiment dataset and the Amazon product review dataset. Particularly on the Amazon product review dataset, DBSCAN is able to efficiently capture users' emotional tendencies and purchasing preferences, and classify users into different groups depending on their ratings and review contents, so offering a strong support for the subsequent value analysis and personalised recommendation.

5 Conclusions

This paper tackles the issue of user value analysis in the social network environment and suggests a hybrid method based on topic model and cluster analysis, which successfully classifies the user values of social networks and e-commerce platforms by means of the DBSCAN clustering algorithm. The paper initially presents the backdrop of social networks and their study of user values; then it discusses the use of topic models and cluster analysis techniques, particularly LDA and DBSCAN, and suggests a novel framework integrating topic models and cluster analysis. Through the extraction of possible themes in social networks and subsequent cluster analysis, the framework is able to detect users' emotive tendencies and values.

Still, this study has several limits. First of all, the DBSCAN clustering method is sensitive to parameter selection, particularly the choice of the two parameters, Epsilon and MinPts, which could influence the clustering outcomes. Different datasets may call for different parameter configurations; hence, the process of parameter tuning calls for several experiments and computations, which could cause instability of the clusterering findings. Second, the quality of text pre-processing and feature extraction determines much how effective and successful topic models – such as LDA – are. While traditional

approaches such TF-IDF were applied for feature extraction in this work, more sophisticated natural language processing tools such pre-trained language models like BERT could be required to improve the efficacy of topic extraction on some complicated datasets.

The following areas will help to improve and broaden future studies:

1 Explore the combination and optimisation of multiple clustering algorithms. This work made use of the DBSCAN clustering method, which although performs well on some datasets has still a difficulty in sensitivity to parameters. We should take into account merging DBSCAN with other clustering techniques, (e.g., K-means, hierarchical clustering, etc.), and applying an integrated learning strategy to combine several clustering outcomes to enhance classification in the future.

2 Multimodal data fusion. Although user behaviour on social networks and e-commerce platforms frequently comprises data in several modalities, (e.g., photos, videos, audio, etc.), this study is essentially dependent on text data for analysis. To build a more complete and accurate analysis model, we can aim to mix multimodal data – e.g., photos, user comments, and behavioural logs – into value analysis going forward.

3 Long-term analysis considering dynamic changes in user behaviour. While in practical applications user values and behaviours often fluctuate over time, this study is focused on stationary datasets for analysis. To observe the evolving trends of user values and behaviours, future studies can take into account combining time series analysis and dynamic clustering techniques, therefore offering real-time analysis for tailored suggestion and precision marketing.

By means of the extension of these future research directions, researchers can enhance the accuracy and applicability of the models, so promoting the development of user values analysis techniques, and so offer more complete and accurate user insights for social networks, e-commerce platforms, and other sectors.

Declarations

All authors declare that they have no conflicts of interest.

References

Abkenar, S.B., Kashani, M.H., Mahdipour, E. et al. (2021) 'Big data analytics meets social media: a systematic review of techniques, open issues, and future directions', *Telematics and Informatics*, Vol. 57, p.101517.

Benlian, A., Titah, R. and Hess, T. (2012) 'Differential effects of provider recommendations and consumer reviews in e-commerce transactions: an experimental study', *Journal of Management Information Systems*, Vol. 29, No. 1, pp.237–272.

Christakis, N.A. and Fowler, J.H. (2013) 'Social contagion theory: examining dynamic social networks and human behavior', *Statistics in Medicine*, Vol. 32, No. 4, pp.556–577.

Egger, R. and Yu, J. (2022) 'A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts', *Frontiers in Sociology*, Vol. 7, p.886498.

Hagen, L. (2018) 'Content analysis of e-petitions with topic modeling: how to train and evaluate LDA models?', *Information Processing & Management*, Vol. 54, No. 6, pp.1292–1307.

Huang, H., Zavareh, A.A. and Mustafa, M.B. (2023) 'Sentiment analysis in e-commerce platforms: a review of current techniques and future directions', *IEEE Access*, Vol. 11, pp.90367–90382.

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.

Kaligotla, C., Yücesan, E. and Chick, S.E. (2022) 'Diffusion of competing rumours on social media', *Journal of Simulation*, Vol. 16, No. 3, pp.230–250.

Kumar, K.M. and Reddy, A.R.M. (2017) 'An efficient k-means clustering filtering algorithm using density based initial cluster centers', *Information Sciences*, Vol. 418, pp.286–301.

Mata, F.J. and Quesada, A. (2014) 'Web 2.0, social networks and e-commerce as marketing tools', *Journal of Theoretical and Applied Electronic Commerce Research*, Vol. 9, No. 1, pp.56–69.

Mittal, M., Goyal, L.M., Hemanth, D.J. et al. (2019) 'Clustering approaches for high-dimensional databases: a review', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 9, No. 3, p.e1300.

Monalisa, S. and Kurnia, F. (2019) 'Analysis of DBSCAN and K-means algorithm for evaluating outlier on RFM model of customer behaviour', *Telkomnika (Telecommunication Computing Electronics and Control)*, Vol. 17, No. 1, pp.110–117.

Newman, M.E. and Park, J. (2003) 'Why social networks are different from other types of networks', *Physical Review E*, Vol. 68, No. 3, p.036122.

Popat, S.K. and Emmanuel, M. (2014) 'Review and comparative study of clustering techniques', *International Journal of Computer Science and Information Technologies*, Vol. 5, No. 1, pp.805–812.

Yang, L., Li, Y., Wang, J. et al. (2020) 'Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning', *IEEE Access*, Vol. 8, pp.23522–23530.

Yue, L., Chen, W., Li, X. et al. (2019) 'A survey of sentiment analysis in social media', *Knowledge and Information Systems*, Vol. 60, pp.617–663.