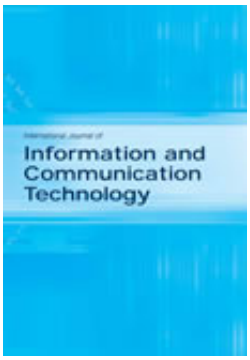


International Journal of Information and Communication Technology

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

Ideological opinion clustering algorithm based on GTE text vector model with inverted index

HuiJie Chai, Weifeng Cai

DOI: 10.1504/IJICT.2025.10072946

Article History:

Received:	26 June 2025
Last revised:	17 July 2025
Accepted:	18 July 2025
Published online:	08 September 2025

Ideological opinion clustering algorithm based on GTE text vector model with inverted index

HuiJie Chai

Faculty of Education,
Shaanxi Normal University,
Xi'an, 710062, China
and
Department of Law,
Henan Police College,
Zhengzhou, 450046, China
Email: jipincai@126.com

Weifeng Cai*

School of Computer and Information Engineering,
Henan University of Economics and Law,
Zhengzhou, 450046, China
Email: chj22567835@126.com
*Corresponding author

Abstract: The study of ideological public opinion on social media platforms has become a crucial focus in public administration and other domains as the internet develops rapidly. Conventional approaches of opinion analysis often suffer with low efficiency of large-scale data processing, inadequate semantic understanding of text, and noise interference. This work therefore suggests an ideological public opinion clustering method (GTE-ICA) grounded on GTE text vector model with inverted index. To enhance the clustering effect of ideological opinion data, the method combines the effective retrieval properties of inverted index with the strong textual semantic representation capacity of GTE model. Experimental data reveal that GTE-ICA has great noise robustness and performs very well on several conventional evaluation criteria. This work offers a reference guide for next development as well as a quick fix for handling ideological opinion data.

Keywords: ideological opinion analysis; GTE model; inverted index; clustering algorithm.

Reference to this paper should be made as follows: Chai, HJ. and Cai, W. (2025) 'Ideological opinion clustering algorithm based on GTE text vector model with inverted index', *Int. J. Information and Communication Technology*, Vol. 26, No. 32, pp.101–120.

Biographical notes: HuiJie Chai enrolled in the Doctoral Program of Educational Leadership and Management at the School of Education, Shaanxi Normal University in 2023. She currently serves as an Associate Professor at Henan Police College. Her main research areas include educational management and civil law.

Weifeng Cai obtained his Master's degree in Ideological and Political Education from Henan University of Economics and Law in 2010. He currently serves as a Lecturer at the Henan University of Economics and Law. His main research areas include political education and computer science.

1 Introduction

Along with the fast growth of internet technology and the general popularity of mobile terminals, the channels for information acquisition and expression have become more and more varied in recent years, and all kinds of ideological views and emotional tendencies in society have shown hitherto unheard-of fast spread and extensive interaction. Particularly social media, news sources, forums, communities, and other platforms have greatly encouraged the generation and spread of public opinion; ideological opinion has thus become a major determinant of social stability, policy making, public sentiment, even emergency response (Adelaja et al., 2018). In this regard, quick and precise identification and analysis of public opinion data has become a frequent need in public administration, social governance, and commercial monitoring.

Mostly in text form, ideological public opinion data comes from a range of sources including news stories, social media comments, and internet forum entries. The conventional approaches of depending just on human screening or keyword-based methods cannot now satisfy the real needs as these textual data are vast in scale, complicated in content, and vary in expression. An important direction in present public opinion research is how to organise and analyse the text data of ideological public opinion using intelligent technology to find possible public opinion hotspots and identify distinct opinion groups (Chen et al., 2019). Among these, text clustering technology has evolved into one of the main instruments in public opinion research thanks to its benefits in subject extraction and information organisation. Large-scale and messy public opinion data can be split into multiple semantically related groups by effective text clustering, therefore exposing the internal structure of many opinions, attitudes, and emotions and giving a technical basis for risk prevention and decision support.

Even although text clustering technology is developing constantly in recent years, two unresolved issues in practical application remain. On the one hand, given complex semantic expressions and context-dependent ideological public opinion texts, which influence the dependability of clustering results, traditional text representation approaches are sometimes difficult to precisely capture deep-level semantic connections. Conversely, the extent of ideological public opinion data is vast and the current clustering techniques still have to be developed in terms of computational efficiency and system reaction speed, so restricting their capacity of real-time analysis and large-scale application (Brady, 2019).

Aiming at the above problems, this paper presents an opinion clustering algorithm based on the combination of GTE text vector model and inverted index, named GTE-ICA, which can efficiently extract the deeper semantic features of the text, and enhance the semantic expression ability between the texts. It can quickly find and arrange material in vast amounts of text data. The two together greatly enhance the computational efficiency of the clustering process as well as the expression correctness of ideological viewpoint texts.

First, we review the present research status and technological development in the field of ideological and public opinion clustering, then we investigate the GTE-ICA algorithm systematically in this paper analysing the shortcomings of the current methods in handling ideological and public opinion data. After that, GTE-ICA's overall design concept and main technological implementation are explored in detail and several sets of tests are built for performance confirmation. In terms of clustering effect and computing efficiency, the experimental findings suggest that GTE-ICA is better than conventional approaches; moreover, it has strong chances for useful application. Finally, this work summarises the findings of the research, evaluates the shortcomings of the present work, and looks ahead to the future direction of the investigation.

2 Relevant work

2.1 GTE text vector model

Aiming at turning unstructured textual data into numerical representations that computers can use, text vectorisation is one of the key foundations in the field of natural language processing. Early on, text representation mostly depended on statistically-based methods such as bag of word (BoW) and term frequency-inverse document frequency (TF-IDF) approaches, which build high-dimensional sparse vectors to represent text by counting word occurrences (Nafis and Awang, 2021). BoW overlooks word order and contextual semantics, however TF-IDF, while it can somewhat solve the word-frequency-dominated issue, has limited expressive power in the face of a vast corpus and complicated environments and cannot adequately capture the deep semantic information of the text.

The GTE text vector model is exactly a text semantic expression framework for general scenarios proposed in this context. Text semantic embedding model based on deep neural network (DNN) and pre-training ideas have become a research hotspot in recent years, and further improvement of the text expression ability depends on it. By means of a consistent neural network architecture, the GTE model seeks to translate text information of various lengths and types into a single low-dimensional dense vector space. Through a unified neural network structure, the GTE model seeks to map text information of various lengths and types into a single low-dimensional dense vector space to capture rich contextual information and deep semantic features in the text and to overcome the limits of conventional approaches in terms of weak semantic expression ability and poor generality.

Particularly based on the design of deep bidirectional coding structure, the GTE model uses large-scale unsupervised corpus to attain the process of multi-granularity and deep semantic modelling of text via self-supervised learning or comparative learning mechanism (Ma et al., 2022). Let the text collecting be D, then D can be written as:

$$D = \{d_1, d_2, \ldots, d_n\} \tag{1}$$

The GTE model achieves the following expression using the mapping function $f(\cdot)$ where every text d_i has several word sequences:

$$v_i = f(d_i) \tag{2}$$

where v_i is the semantic representation of text d_i in the low-dimensional vector space; k is the set vector dimension. By cosine similarity, Euclidean distance, and other measures, this vector allows one to objectively estimate the semantic similarity between texts, so facilitating the later text clustering, classification, and retrieval activities.

GTE models emphasise more adaptability and flexibility of the expression framework than conventional pre-trained language models such as BERT, RoBERTa, S-BERT, etc.; they thus can flexibly change the model structure and training strategy depending on the needs of the application, so considering the expression impact and computational efficiency (Singh and Mahmood, 2021). To guarantee the efficient capture of long-distance dependencies and complicated semantic structures in text, GTE models also typically mix positional coding, multi-layer semantic extraction and global information aggregation algorithms.

Particularly in large-scale text data modelling and multi-language, multi-domain scenarios show good adaptability and scalability. GTE text vector models have thus been extensively applied in a variety of text processing tasks, including text classification, semantic matching, information retrieval, text clustering and sentiment analysis, etc., currently. The expression ability and application scope of GTE models are expected to be further expanded with the continuous development of pre-training technology, multimodal fusion and lightweight model design, so becoming a significant technical direction in natural language processing and text semantic modelling.

In terms of in-depth semantic expression, contextual information retention and large-scale text adaptation ability, GTE text vector models are overall better than conventional BoW models TF-IDF methods and early word vector models. They have thus become one of the main development trends in the field of text data modelling.

2.2 Inverted index

Widely employed in search engines, information retrieval and large-scale text analysis systems, inverted index is a classic and effective data structure for text retrieval. The fundamental concept is to create a direct mapping relationship between every keyword in the text collection and the list of documents including the keyword, so avoiding the high computational costs connected with the conventional sequential scanning of all the documents and greatly increasing the retrieval efficiency. Inverted index by the building of the 'keyword → document collection' mapping, to achieve a fast positioning of the target document, especially appropriate for handling enormous text data scenarios.

Inverted index mostly consists of a dictionary and posting list. Often using hash tables, B-trees or prefix trees and other data structures to lower the complexity of the lookup time, dictionary maintains all the independent words in the text collection and facilitates effective lookup operations (Xing et al., 2025). The inverted table, which records all the document id including the word, as well as optional additional information such as the word frequency (TF), the position of the word in the document (position index), etc., is a chained table or array matching every word in the dictionary. These further details support the later importance of the word. This extra data gives great help for later relevance computation and improved retrieval.

Three main phases comprise the construction of an inverted index: text preprocessing, word item extraction and index generating. Usually aiming to improve the degree of normalisation of lexical items and retrieval quality, text preparation consists in procedures including word splitting, elimination of deactivated words, stemming

reduction, etc. The fundamental unit of indexing is found via lexical item extraction; the granularity directly influences the space overhead and indexing accuracy. Usually employing compression technology to conserve storage space and increase reading efficiency, index generation is to write the processed lexical items and accompanying document information into inverted index structure.

In actual applications, inverted index not only helps the fast placing of keywords but also efficiently assists text clustering, opinion analysis, and other difficult text processing chores. More fine-grained similarity measure and context analysis can also be accomplished depending on the information of word frequency and position to enhance the accuracy of clustering results. For instance, inverted index can rapidly identify a collection of text including the same keywords in the text clustering process, so lowering the number of candidate pairs for computing similarity and greatly lowering the computational complexity of the clustering method.

Furthermore, supporting effective incremental update and distributed storage techniques for big-scale dynamic text libraries is inverted index. Ensuring real-time retrieval results, the incremental update technique lets the system update the inverted structure when new documents are added or old documents are altered (Corley et al., 2018). Parallel computing and distributed storage help to satisfy internet-level text data processing requirements and expand the inverted index to huge data environments.

Algorithm 1 presents a condensed pseudocode model of the inverted index's building and search:

Algorithm 1 Pseudocode for constructing and querying inverted index

```
Input: Document set, query keywords.
Output: Inverted index, query results.
begin
Initialise empty inverted index
for each document in document set do
    Preprocess document (tokenise, remove stop words, stemming)
    for each token in document do
        if token not in inverted index then
            Initialise posting list for token
        end if
        Add current document ID to token's posting list
    end for
end for
Initialise empty result set
for each keyword in query do
    if keyword exists in inverted index then
        Retrieve posting list for keyword
        Add documents from posting list to result set
    end if
end for
Optionally, rank documents in result set based on term frequency or other metrics
```

20: return inverted index and result set
21: **end**

Finally, inverted index has evolved as the fundamental basis of large-scale text information retrieval and analysis due to its effective query performance and succinct framework. Its fast screening of text, reduction of computational redundancy and support of real-time updates give a strong technical guarantee for text-based clustering algorithms and ideological opinion analyses; it is also a necessary key component for obtaining efficient opinion data processing.

2.3 Ideological opinion clustering algorithm

An important direction in the field of natural language processing, ideological opinion clustering algorithms seek to extract text collections with comparable viewpoints, sentiments, or subjects from a vast amount of ideological opinion text data, and subsequently efficiently mine and analyse them. Mass ideological opinion data is fast generated with the broad use of social media, news platforms and forums; so, how to find possible opinion trends and perspective shifts in this enormous textual data becomes a critical issue in academics and business.

Early ideological opinion clustering approaches mostly derived from conventional text clustering techniques, including K-means, hierarchical clustering, and unsupervised learning algorithms such as DBSCAN (Wang, 2025). These approaches frequently reveal certain restrictions when dealing with complicated public opinion clustering even if they are somewhat straightforward and easy to apply. For instance, the K-means algorithm depends on the Euclidean distance metric, which fails to adequately depict the intricate semantic linkages between texts, and is particularly useless when confronting ideological public opinion with many expressions and complicated emotional colours. Although hierarchical clustering can produce more exact clustering results, its computing expense is significant, particularly in cases with a lot of data and low clustering efficiency.

Text clustering techniques based on vector space models have progressively taken front stage as text representation techniques improve. Although they can handle some basic clustering chores more effectively, TF-IDF and BoW ignore the contextual and semantic relationships of words in the text, and it is challenging to deal with ideological and opinionated data that include complicated semantics. TF-IDF and BoW represent features by counting the frequency of words in the text. Researchers have started using deep learning-based text representations to address this challenge (Oyewole and Thopil, 2023). By using contextual co-occurrence relations between words and mapping words in text into a low-dimensional dense vector space, word vector models such as Word2Vec, GloVe, and FastText can efficiently maintain semantic information. These techniques, however, still mostly concentrate on the word level and cannot fairly depict the whole semantics of phrases or extended texts.

DNN-based text representation techniques have progressively become a research hotspot in recent years to improve the impact of ideological opinion clustering even further. Particularly, by means of many unsupervised corpora training, text embedding methods based on pre-trained language models (e.g., BERT, GPT, etc.) can capture deep semantic information and contextual dependencies in the text, so improving the quality of text representation. Ideological opinion clustering's precision and potency are much enhanced in this approach. For instance, the BERT model generates text vectors with

strong semantic information by concurrently understanding the backward and forward relationships in the context using a bidirectional encoder (Eke et al., 2021). These pre-training-based deep learning models help clustering algorithms not only to find textual similarities but also to better grasp the emotions and themes suggested in the text.

Apart from conventional clustering techniques, various algorithms derived from graph neural network (GNN) have been progressively included into the study of ideological public opinion in recent years. To attain more accurate text grouping, GNNs employ the nodes and edges in the graph for the propagation of information and describe the intricate structural links between texts using this ability (Yang et al., 2021). Using approaches such as graph convolutional network (GCN) for feature learning and interpreting ideological opinion texts as graph structures with nodes signifying texts and edges denoting comparable links between texts, researchers have enhanced the effectiveness of clustering algorithms.

Furthermore, progressively becoming a new avenue of research are clustering techniques grounded in sentiment analysis and topic modelling. Often in ideological opinion analysis, sentiment tendency is a crucial component in determining several opinion subjects. Combining sentimental data with conventional text clustering might thus improve the accuracy and application value of clustering results even more. Some studies, for instance, have suggested combining sentiment labels and theme models, which not only take into account the similarity of text content in the clustering process but also integrate sentiment tendency, so improving the clustering results in reflection of the emotional changes and social attitudes of public opinion.

Though several techniques have shown amazing success in ideological opinion grouping, they nevertheless have certain difficulties. First, the sentiment of ideological opinion texts is complicated and fluctuating; hence, distinct texts may have complicated implicit sentiment information and polysemy, which increases the demand on clustering techniques (Memon et al., 2021). Second, the distribution of ideological opinion data is sometimes unequal and the amount of opinion texts on some popular themes far surpasses that on other topics, which results in the fact that the clustering results may be biassed by popular topics, therefore influencing the general clustering impact. Furthermore, the real-time and dynamic character of text data is another major obstacle confronting ideological public opinion clustering at present and how to perform efficient real-time clustering in fast changing data remains an urgent issue as public opinion events are continuously developing.

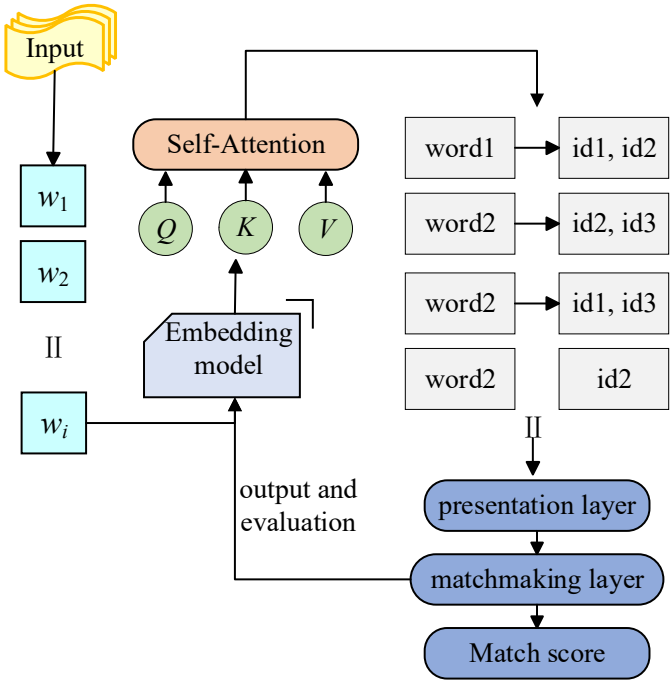
From conventional clustering approaches to contemporary techniques based on deep learning and graph neural networks, ideological opinion clustering algorithms have made notable development overall. Future research will thus concentrate more on how to mix techniques such as sentiment analysis, topic modelling, multimodal data and real-time updating to cope with the diversity and dynamics in the analysis of ideological public opinion, so as to achieve more efficient and accurate opinion clustering, given the expanding size of data and the growing complexity of opinion content.

3 Algorithmic framework for ideological opinion clustering

This research suggests a transparent framework GTE-ICA to apply the ideological opinion clustering method based on GTE text vector model and inverted index, see

Figure 1. The text preprocessing layer first cleans and normalises the original text; then, the GTE text vectorising layer converts the text into dense vectors and extracts the semantic information; subsequently, the inverted index construction layer improves the retrieval efficiency by building an inverted index; in the similarity calculation and clustering layer, clustering is performed based on text similarity; finally, the result output and evaluation layer analyses the clustering effect through evaluation and visualisation. By use of the cooperative effort of these five modules, the GTE-ICA method may effectively finish the ideological opinion clustering process.

Figure 1 Framework of the GTE-ICA algorithm



3.1 *Text pre-processing layer*

Aimed to offer high-quality input for the clustering analysis of ideological opinion data, this layer is the first in the GTE-ICA method. Effective preprocessing of the text can help to increase the accuracy of the later analysis since ideological opinion writings typically include a lot of noise, redundant information, and non-semantic high-frequency terms. First, the segmentation procedure divides the unprocessed text into distinct lexical units, therefore offering the fundamental building block for more work. Deactivation then eliminates words that appear more frequently and have less semantically, such as ‘the’ and ‘is’, so lowering noise and enhancing the text feature differentiation.

Usually, stemming is not relevant in the processing of ideological opinion data since such texts include many distinct words with emotional colours and opinions; so, the preservation of word form changes helps the algorithm better capture the emotional and thematic variations of the text. Consequently, greater focus should be on preserving lexical variety and accuracy than on stemming.

Furthermore, text standardisation activities like lowering all characters to lowercase and eliminating punctuation aid to unite the text format so that texts from several sources could be compared on the same dimension and guarantees data consistency. These preparation techniques help to remove interfering information and preserve important information in the text, therefore offering clean and significant input for next text vectorisation and clustering.

The TF-IDF approach is applied to improve the quality of text vectorisation by assessing the relevance of words in a document, thereby strengthening the textual representation (Abubakar et al., 2022). TF-IDF's formula is presented below:

$$\text{TF-IDF}(t, d) = TF(t, d) \times IDF(t) \tag{3}$$

where TF(t, d) is the word frequency of lexical item t in document d; IDF(t) is the inverse document frequency of lexical item t across all the documents. TF-IDF not only reduces highly frequent undesired words but also increases the weight of significant words in each document. This procedure guarantees exact features for the next cluster analysis of the text, helps to extract representative information, and avoids the interference of irrelevant content.

By means of these extensive preprocessing stages, the text data is cleaned and normalised, so preserving the important information and sentiment tendencies in them, thus laying a strong basis for further GTE model vectorisation and clustering analysis.

3.2 GTE text vector layer

The primary responsibility of this layer in the GTE-ICA method is to convert the preprocessed text into an excellent dense vector representation for the next cluster analysis. Using a vast-scale corpus, the GTE model builds a DNN to learn about the global semantic aspects of text. First in this layer the text input goes through the word embedding process, which maps every word into a low-dimensional dense vector. Semantically comparable words will thus be closer together in the vector space, hence improving the capacity of the model to represent semantic links.

Using multi-layer neural network architecture, the GTE model performs text vectorisation where the input text is layer-by-layer feature extracted via a multi-layer encoder (Minaee et al., 2021). With great expressive and discriminative ability, the last produced text vectors can capture both local lexical meaning and global contextual information by means of the deep learning of these layers.

Imagine an input text T:

$$T = \{w_1, w_2, \ldots, w_n\} \tag{4}$$

where every word w_i maps to a d-dimensional word vector v_{w_i}. The individual word vectors taken together produce the final vector representation V_T of the text. Formula for text vectorisation is:

$$V_T = \sum_{i=1}^{n} \alpha_i v_{w_i} \tag{5}$$

where v_{w_i} is the vector form of the word w_i and α_i is the weight of every word vector; typically, these values are changed in line with word frequency or another approach.

The GTE model uses the self-attention mechanism to increase the effect of text vectorisation therefore strengthening the representation capacity of the model. By computing the correlation between individual words in the text, the self-attention mechanism dynamically changes the weight between word vectors, therefore enhancing the capacity to grasp significant information (Zheng et al., 2020). The self-attention mechanism's formula is:

$$\text{Attention}(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where $\sqrt{d_k}$ is a constant used for scaling to minimise numerical instability in high-dimensional spaces, Q, K and V denote the query, key and value matrices accordingly. This allows the GTE model to dynamically allocate varying weights to every word, therefore more precisely capturing the semantic links in the text.

By means of these deep learning methods, the GTE model produces more accurate and richer representations of text vectors, so enhancing text difference and semantic expression in ideological opinion clustering analysis. Subsequent cluster studies will build on these text vectors, therefore exposing the underlying themes and emotional inclinations in ideological public opinion.

3.3 Reverse index building layer

Especially adapted for the fast processing of vast-scale ideological opinion data, the inverted index creation layer is crucial in the GTE-ICA algorithm to improve the retrieval efficiency and speed the similarity computation. Large amounts of text and redundant information in ideological and public opinion data mean that if large-scale two-by-two matching is done straight based on text vectors, the computing overhead is great, and the efficiency is low. By creating the mapping link between keywords and text numbers, inverted index accomplishes effective text localisation and screening to solve this problem, so considerably improving the pace of general cluster analysis.

Inverted index is essentially a means of recording, for every keyword, its text number or position in the document collection. Particularly, let the text collection be $\{D_1, D_2, \ldots, D_m\}$, the keyword collection be $\{t_1, t_2, \ldots, t_m\}$, and the inverted index's fundamental form be written as:

$$t_i \to \left\{D_j \middle| t_i \in D_j\right\} \tag{7}$$

where t_i marks the i^{th} keyword. By means of this structure, the GTE-ICA algorithm may rapidly recover all the texts including a given keyword, therefore avoiding repetitive traversal of the whole text data and optimising the retrieval and screening effectiveness.

In the processing of ideological opinion data, the choice of keywords is very important to guarantee the efficacy of inverted index. The TF-IDF approach combined with the weighted and screened keywords in the text guarantees that the chosen keywords are both representative and efficient in differentiating various texts (Qaiser and Ali, 2018). The computation of keyword weights follows this formula:

$$\text{Weight}(t_i) = \text{TF-IDF}(t_i, D_j) \times \delta(t_i) \tag{8}$$

where TF-IDF(t_i, D_j) indicates the TF-IDF weight of the lexical item t_i in the document D_j, $\delta(t_i)$ is the domain-defined adjustment coefficient used to dynamically change the relevance of the keywords based on the traits of the real opinion corpus. By use of this weighting technique, the inverted index can more precisely represent the central content of the text and minimise the interference with erroneous information.

The GTE-ICA method not only enables effective keyword retrieval but also helps with the subsequent similarity coarse screening and text grouping, thereby greatly lowering the total computing cost of the system with the inverted index structure. In large-scale ideological opinion data analysis, the introduction of this technology essentially assures the efficiency and practicality of the algorithm.

3.4 Similarity calculation and clustering layer

The core of the GTE-ICA method to attain ideological opinion clustering is this layer. Following text preprocessing, vectorisation and inverted index construction, the system must precisely assess the degree of semantic association between various texts by means of similarity computation, so grouping the texts with similar content and related topics into the same category and so exposing the structural aspects and possible topics in the opinion data.

In this layer, the semantic relationship between several texts is initially assessed using the similarity measure between the dense text vectors produced by the GTE text vectorisation layer (Tian et al., 2023). Calculated as cosine similarity using the following formula, the text similarity:

$$\text{Sim}(V_i, V_j) = \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|} \tag{9}$$

where the symbol · is the inner product of the vectors; $||V_i||$ is the vector modulus when V_i and V_j represent the dense vector representations of text i and text j, respectively. The cosine similarity reflects the proximity of the texts in the high-dimensional semantic space rather well; so, the contents of the texts are more similar the higher their value.

Following the text similarity computation, the GTE-ICA algorithm presents the density-based DBSCAN clustering method for the features of large scale of ideological public opinion data, complicated distribution of categories, and significant noise interference. Appropriate for the practical needs of the uncertainty of the number of topics in ideological opinion data, the method automatically identifies areas of different densities by analysing the local density relationships of data points, so forming a clustering structure and eliminates the need of specifying the number of categories in advance.

DBSCAN first defines two main parameters: the minimal number of neighbourhood samples *MinPts* and the radius of the neighbourhood ε (Li, 2020). Any data point p has a neighbourhood defined as:

$$N_\varepsilon(p) = \{q \mid \text{dist}(p, q) \le \varepsilon\} \tag{10}$$

where dist(p, q) is the distance between data points p and q; $N_\varepsilon(p)$ is the set of all the neighbourhood points within the radius ε. After that, the cluster structure is developed recursively to create a whole one. Data point p is regarded as a core point when the total number of points in the neighbourhood exceeds *MinPts*; this is then expanded recursively to generate a whole cluster structure.

The GTE-ICA algorithm can precisely identify various thematic groups and possible opinion trends in large-scale ideological opinion data by means of the combination of similarity computation and DBSCAN clustering, so enhancing the clustering effect and the general analytical capacity of the system.

3.5 Results output and evaluation layer

The main component of the GTE-ICA algorithm is the result output and evaluation layer; it seeks to present and analyse the clustering results in a comprehensive manner, therefore offering a trustworthy basis for later processing of ideological opinion clustering data. Following the earlier stages of text preprocessing, vectorisation, inverted index building and similarity computation, GTE-ICA can effectively divide ideological public opinion text data into several subject categories. By means of the collation of clustering findings, the result output layer shows the text content and representative properties of every cluster, therefore enabling later analysis and interpretation.

The assessment layer contrasts the clustering results with the real labels (e.g., manually labelled topics) to find possible mistakes or deviations in the clustering process, therefore enabling a full evaluation of the clustering impact. To this aim, the text-based mutual information metric measures the consistency between the clustering findings and the actual labels (Chaudhary et al., 2019). The mutual information computation yields:

$$I(C, L) = \sum_{c \in C} \sum_{l \in L} p(c, l) \log \frac{p(c, l)}{p(c)\, p(l)} \tag{11}$$

where C is the set of clustering results; L is the set of real labels; $p(c, l)$ is the probability that clustering c and label l occur simultaneously; $p(c)$ and $p(l)$ respectively indicate the marginal probability of clustering c and label l, respectively. Calculating the mutual information helps one to evaluate the relationship between the actual labels and the clustering results; a higher number denotes more consistent clustering results with the real labels, thereby indicating a better clustering effect.

Furthermore, important during the evaluation process is the stability and adaptability of the method to guarantee that the GTE-ICA algorithm may produce consistent and high-quality clustering results under many datasets and situations. By constantly optimising the clustering parameters, algorithm design, and processing flow, the evaluation layer increases the generalisation capacity and resilience of the algorithm in complicated ideological opinion clustering data.

By means of extensive result output and evaluation, the GTE-ICA method may efficiently expose possible themes, emotional trends and social hotspots in ideological public opinion, so offering useful reference and support for public opinion monitoring, trend analysis and decision making.

4 Experimental design and results

4.1 Experimental data

Derived from Sina Weibo and comprising a significant volume of social media text data, the dataset used in this experiment is Weibo-Public-Opinion-Analysis. The collection consists of Weibo text, user interaction records, and sentiment labels. Particularly useful in ideological opinion clustering, the dataset is fit for tasks of opinion analysis, sentiment analysis, and event detection.

Table 1 provides the specifics of this dataset.

Table 1 Dataset information for Weibo-Public-Opinion-Analysis

Dataset name	Weibo-Public-Opinion-Analysis
Data source	Sina Weibo
Data size	Contains multiple Weibo posts, the exact size varies by version
Data content	Weibo text, user information, number of comments, retweets, likes, sentiment labels
Applicable fields	Sentiment analysis, public opinion analysis, event detection, topic modelling
Additional notes	The dataset includes public data from multiple topics and events on social media, suitable for large-scale sentiment analysis and public opinion research

With this dataset, the GTE-ICA method can examine public opinion dynamics, sentiment trends and topic clustering on Weibo, thereby supporting later opinion monitoring and event detection. The size and variety of the dataset make it a significant reference in the analysis of ideological public opinion, which may help researchers detect public mood and hot issues and thus advance the field of public opinion research development.

4.2 Experimental evaluation

Two often used normalisation metrics are employed in this experiment to assess the performance of the GTE-ICA algorithm in ideological opinion clustering: the normalised mutual information (NMI) and the normalised adjusted rand index (NARI). Both the evaluation of clustering results and the effective quantification of the clustering quality of the method depend on these two criteria.

The similarity between the clustering findings and the true labels is measured by NMI, which ranges from 0 to 1 with 1 denoting total agreement and 0 denoting total irrelevance (Yuan et al., 2021). In tests to assess the applicability of clustering results to real labels, mutual information has been extensively applied as an evaluation criterion. NMI standardises the raw mutual information so that its value is no longer distorted by the number of categories and can more fairly represent the quality of the various number of clusters, therefore eliminating the effect brought about by the varying counts. The recipe is as follows:

$$NMI = \frac{I(C, L)}{\sqrt{H(C) \cdot H(L)}} \tag{12}$$

where $H(C)$ and $H(L)$ are the entropy of the clustering result C and the true label L correspondingly; $I(C, L)$ is the mutual information between them. Normalisation of this formula allows NMI to be adjusted to various clustering challenges and becomes more universal.

Furthermore, NARI is a crucial assessment tool for comparing genuine labels with clustering outcomes (Jeon et al., 2023). NARI can remove the bias resulting from the variation in the number of categories, therefore offering a more realistic assessment. The mixture is:

$$NARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \tag{13}$$

where RI is the rand index, which gauges the consistency of clustering of pairs of data points; $E[RI]$ is the expected value, which indicates the RI in case of irrelevant clustering results; $\max(RI)$ is the highest value of the RI; usually ranging from –1 to 1, the values of the normalised adjusted RI reflect 1 for perfect consistency, 0 for the same as a random outcome, and a negative number signifying worse than a random result.

These two normalisation criteria allow an objective evaluation of the GTE-ICA algorithm's performance in ideological opinion clustering. Apart from these automatic assessment criteria, manual review is also crucial for guaranteeing the efficiency of the algorithm. Researchers will personally examine the thematic relevance and sentiment consistency of every cluster by hand reviewing the clustering results in the manual assessment process. Furthermore, validating the accuracy of the algorithm and offering more understandable input for optimising the algorithm and enhancing the clustering results is the manual study of the clustering results.

4.3 Horizontal algorithm comparison

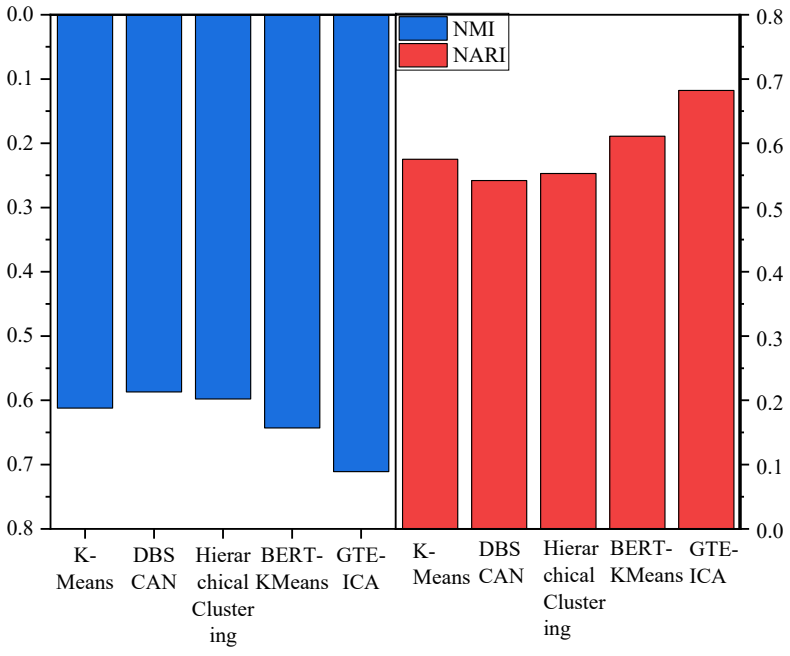
This experiment is intended as a comparative experiment to confirm the advantages of GTE-ICA in clustering accuracy and stability by means of multiple mainstream clustering techniques, so enabling a thorough investigation of the general performance of the GTE-ICA algorithm in the task of ideological opinion clustering. Aiming to give strong support for the efficacy of the ideological opinion clustering algorithm, the experiment is based on the same opinion dataset and a consistent experimental technique to ensure that the test findings of various methodologies are similar.

This experiment chooses K-means, DBSCAN, hierarchical clustering and BERT-KMeans. Classical text clustering algorithms are K-means and DBSCAN; hierarchical clustering reflects the hierarchical structure of the data; BERT-KMeans aggregates pre-trained text and vector models, therefore somewhat enhancing the clustering performance. Vector model that in some measure enhances text representation. These techniques have a wide application base and reflect several kinds of clustering concepts, therefore enabling a thorough comparison and analysis of the consequences of GTE-ICA.

To assess the consistency and similarity between the clustering outcomes and the true labels, respectively, the trials used NMI and NARI as evaluation measures. Every group of tests is repeated numerous times; the average value is obtained as the last outcome to guarantee the dependability and stability of the conclusion.

Figure 2 exhibits the experimental results.

Figure 2 Experimental results of algorithm comparison (see online version for colours)



First, the GTE-ICA method performs really well on all evaluation criteria, especially on the two normalisation measures, NMI and NARI, which are noticeably better than other methods. Compared to other clustering techniques, particularly in relation to BERT-KMeans (0.643), which improves by roughly 6.8%, GTE-ICA earns a high score of 0.711 on the NMI metric. This result reveals that GTE-ICA is more able to detect the underlying subject structure and sentiment groups in microblog text and more exact in its ideological opinion clustering effect on ideological opinion data.

By comparison, the conventional K-means and DBSCAN methods work less effectively. Although the NMI of K-means is 0.612 and that of DBSCAN is 0.587, these two algorithms fail to make appropriate use of the semantic content in textual data, therefore producing less thorough and accurate clustering results. While DBSCAN, which is not able to adapt to the high-dimensional and sparse character of textual data, also fails to provide better results in the clustering of opinion data, K-means may be affected by the choice of the initial centroids, which leads to more fluctuations in the clustering results. DBSCAN also lacks in significant enough clustering findings in opinion data.

Hierarchical clustering is usually applicable to small-scale datasets but may not be suitable for large-scale social media data due to high computational complexity or insufficiently fine-grained hierarchical division, even if it can better reflect the hierarchical relationship of data. Hierarchical clustering is usually applicable to small-scale datasets. Either the hierarchical division or great computational complexity is insufficient to produce appropriate clustering results.

Combining the strong text representation capabilities of BERT, which gets rather decent results with NMI and NARI values of 0.643 and 0.611, respectively, BERT-KMeans achieves comparatively good results as well as standard clustering techniques.

BERT-KMeans fails to show the benefits of GTE-ICA, most likely due to the textual representation of BERT not being fully integrated into the optimisation process of the clustering algorithm, compared with GTE-ICA. This especially relates to complex emotions and diverse topics in opinion data.

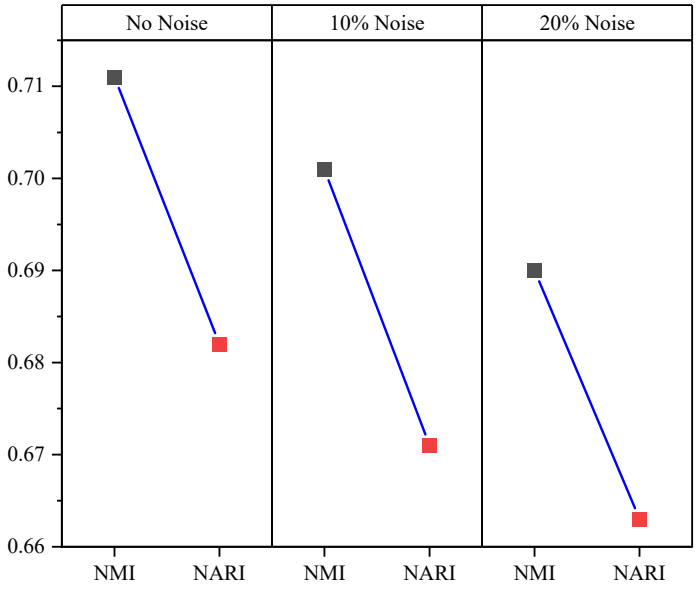
In this experiment, the GTE-ICA method shows superior accuracy and stability in the ideological opinion clustering task, so it greatly beats the other compared methods. This implies that GTE-ICA has more benefits for activities like opinion monitoring and social topic analysis and is more suited to record the fundamental structural and sentiment information in microblog articles.

4.4 Vertical robustness analysis

By means of varying noise levels in the dataset, the experiment assesses the performance of the GTE-ICA algorithm in the face of irrelevant data and incomplete information, therefore testing the resilience of the system under diverse noise disturbances. In practical applications, including extraneous material, spam or spelling mistakes, which could compromise the accuracy of the clustering operation, ideological opinion data are often influenced by a range of disturbances. Thus, verifying the dependability of an algorithm in practical applications depends on assessing its performance in noisy surroundings.

The noise levels are designed to replicate several degrees of data disturbance. Three noise levels were chosen for the experiments: initially, in the case of no noise, the dataset was left as it was without including any disruptive information. To replicate modest noise interference, a 10% portion of the dataset was randomly changed at the 10% noise level with garbled information, meaningless text, or noisy vocabulary. Finally, the 20% noise level added more useless information, mimicking a more complicated noise environment. These several noise level settings seek to investigate the GTE-ICA algorithm's adaptability and robustness under noisy environments.

Figure 3 Experimental results of noise interference (see online version for colours)



To fully assess the consistency and comparability between clustering outcomes and genuine labels, the experimental evaluation applied NMI and NARI as measures. Figure 3 shows the experimental results.

The experimental results allow one to investigate in detail the performance of GTE-ICA method under various noise levels. With NMI and NARI of 0.711 and 0.682 respectively, the method produces the best clustering results under no noise condition. This demonstrates great clustering accuracy and reveals that GTE-ICA can investigate the possible themes and sentiment groupings in ideological opinion data completely. The method can partition the data exactly without noise, and the similarity between the clustering findings and the genuine labels is strong.

The NMI and NARI values exhibit varied degrees of decline with the noise level, but the change is minor. About 1.4% and 1.6% lower than the no-noise case, respectively, NMI falls to 0.701 in the 10% noise situation while NARI falls to 0.671. This indicates that, when modest noise is included in the data, GTE-ICA still preserves the accuracy and consistency of clustering well. The general clustering effect is good even if noise affects the performance of the method; this indicates that the algorithm has great resilience under noise interference.

Under 20% noise, the NMI and NARI values drop even more to 0.690 and 0.663, respectively, with a concomitant loss of around 2.9% and 2.8%). This suggests that although clustering may still be carried out efficiently, the accuracy and consistency of clustering somewhat diminish compared to the no noise and 10% noise situations when the noise percentage reaches 20%. The clustering impact of GTE-ICA is compromised to some amount. The general decline is not noteworthy, nevertheless, suggesting that GTE-ICA has excellent anti-interference capacity and can handle more noisy data.

Collectively, the GTE-ICA method exhibits great adaptability and stability in the face of varying noise intensity. The method is still able to retain a high degree of accuracy and consistency in noisy surroundings despite a little reduction in clustering effect. GTE-ICA is still able to efficiently extract the valuable information in the text and guarantee the clustering quality especially in cases of a larger proportion of noise. This gives great support for the general usability of the algorithm in actual applications, particularly in the monitoring of public opinion with large proportion of noisy data, GTE-ICA can exert its great robustness and stability.

Following the quantitative analysis, manual study can help to confirm the validity of the clustering results so enabling more extensive evaluation of the GTE-ICA technique. Apart from verifying the thematic consistency of the clustering results of the algorithm, manual analysis also evaluates the reasonableness and interpretability of clustering under many noise levels. More intuitively the performance of the method in practical applications may be judged by the hand evaluation of the clustering results, particularly about how to manage noise interference.

Table 2 displays the scoring results of the manual analysis in which the quality of clustering is manually assessed depending on the clustering outcomes under various noise levels. Mostly, the scoring criteria are assessed in relation to the dimensions of consistency of text inside every cluster, variation among clusters, and interpretability of themes.

Table 2 Manual evaluation scores for clustering results under different noise levels

Noise level	*Manual evaluation score (out of 10)*
No noise	8.5
10% noise	8.1
20% noise	7.6

The manual analysis' score findings reveal that the quantity of noise lowers the rationality of the clustering results. In the absence of noise, the obtained clustering results had the highest manual score of 8.5, meaning that they performed rather well in terms of thematic consistency and cluster differentiation. Regarding the 10% noise, the score dropped somewhat to 8.1; although the noise created some interference with the clustering findings, the general clustering results were still adequate. In the case of 20% noise, the manual score falls to 7.6, suggesting that the effect of noise on the clustering effect is more significant and the thematic consistency and differentiation of the clustering is lowered, but it is still able to identify several groups of public opinion overall.

These hand scoring results support even further the resilience of GTE-ICA in the face of various noise settings, particularly in the situation of significant data interference; the method nevertheless produces more accurate and interpretable clustering results.

5 Conclusions

5.1 Summary of the study

This work proposes an ideological opinion clustering algorithm (GTE-ICA) grounded on the GTE text vector model with inverted index using inverted index. GTE-ICA effectively extracts possible topic and sentiment information in thought-opinion data by combining the textual representation capabilities of the GTE model with the efficient retrieval properties of inverted index. In terms of the two-normalisation metrics, NMI and NARI, the GTE-ICA algorithm greatly beats conventional clustering techniques including K-means, DBSCAN, hierarchical clustering, and BERT-based clustering approaches. Dealing with large-scale datasets, the technique still shows great clustering accuracy and stability, therefore demonstrating its possible use in ideological opinion clustering.

Furthermore, this work confirms, by experiment 2, the strength of the GTE-ICA algorithm. Although the inclusion of noise causes a small decline in the results, GTE-ICA is still able to retain high clustering results at various noise levels; so, the general performance is still superior to other compared techniques. Particularly in chaotic surroundings, where the thematic coherence and distinctiveness of the clustering results remain high, the manual evaluation results also indicate the efficiency of the algorithm in pragmatic uses. These findings show that GTE-ICA has a broad practical application value, good clustering capacity, and can manage complicated data interference in real-world settings as well.

5.2 Limitations and improvements

The GTE-ICA method has certain restrictions even if it performs well in the ideological opinion grouping challenge. First, in complicated texts, especially when dealing with long texts or polysemous words, the GTE text vector model might not be able to adequately represent deep semantic information. Future research can add more advanced text representation methods, including pre-trained language models based on the transformer architecture, which can better handle complicated contexts and polysemous word difficulties and improve the semantic representation of text, so helping to overcome this challenge.

Second, although it has great retrieval efficiency when handling large-scale data, inverted index still lacks the semantic relevance of text. Future study can investigate retrieval strategies including depth, such as strengthening the semantic understanding capacity of inverted index by means of semantic indexing or graph embedding techniques, so improving the efficacy and quality of clustering.

Furthermore, the noise simulation in this work mostly depends on the introduction of somewhat uniform and nature simple random noise. Different applications call for more varied forms of noise in thought-opinion data, such as spelling mistakes, grammatical faults, information redundancy, and emotional ambiguity, which affect the text differently. More sophisticated noise models for various kinds of noise and focused denoising strategies to improve the robustness of the GTE-ICA algorithm in more complicated and realistic data contexts can be designed in future work.

These developments enable the GTE-ICA algorithm to handle more complicated text data in the next studies and raise its dependability and efficiency in practical applications including opinion analysis and event detection.

Declarations

This work is supported by the 2024 Henan Provincial Planning Office Soft Science Project (No. 2024BFX034).

All authors declare that they have no conflicts of interest.

References

Abubakar, H.D., Umar, M. and Bakale, M.A. (2022) ‘Sentiment classification: review of text vectorization methods: bag of words, TF-IDF, Word2vec and Doc2vec’, *SLU Journal of Science and Technology*, Vol. 4, No. 1, pp.27–33.

Adelaja, A.O., Labo, A. and Penar, E. (2018) ‘Public opinion on the root causes of terrorism and objectives of terrorists: a Boko Haram case study’, *Perspectives on Terrorism*, Vol. 12, No. 3, pp.35–49.

Brady, H.E. (2019) ‘The challenge of big data and data science’, *Annual Review of Political Science*, Vol. 22, No. 1, pp.297–323.

Chaudhary, C., Goyal, P., Tuli, S., Banthia, S., Goyal, N. and Chen, Y-P.P. (2019) ‘A novel multimodal clustering framework for images with diverse associated text’, *Multimedia Tools and Applications*, Vol. 78, pp.17623–17652.

Chen, X., Duan, S. and Wang, L-D. (2019) ‘Research on clustering analysis of internet public opinion’, *Cluster Computing*, Vol. 22, pp.5997–6007.

Corley, C.S., Damevski, K. and Kraft, N.A. (2018) 'Changeset-based topic modeling of software repositories', *IEEE Transactions on Software Engineering*, Vol. 46, No. 10, pp.1068–1080.

Eke, C.I., Norman, A.A. and Shuib, L. (2021) 'Context-based feature technique for sarcasm identification in benchmark datasets using deep learning and BERT model', *IEEE Access*, Vol. 9, pp.48501–48518.

Jeon, H., Kuo, Y-H., Aupetit, M., Ma, K-L. and Seo, J. (2023) 'Classes are not clusters: improving label-based evaluation of dimensionality reduction', *IEEE Transactions on Visualization and Computer Graphics*, Vol. 30, No. 1, pp.781–791.

Li, S-S. (2020) 'An improved DBSCAN algorithm based on the neighbor similarity and fast nearest neighbor query', *IEEE Access*, Vol. 8, pp.47468–47476.

Ma, C., Zhang, W.E., Guo, M., Wang, H. and Sheng, Q.Z. (2022) 'Multi-document summarization via deep learning techniques: a survey', *ACM Computing Surveys*, Vol. 55, No. 5, pp.1–37.

Memon, M.Q., Lu, Y., Chen, P., Memon, A., Pathan, M.S. and Zardari, Z.A. (2021) 'An ensemble clustering approach for topic discovery using implicit text segmentation', *Journal of Information Science*, Vol. 47, No. 4, pp.431–457.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M. and Gao, J. (2021) 'Deep learning--based text classification: a comprehensive review', *ACM Computing Surveys (CSUR)*, Vol. 54, No. 3, pp.1–40.

Nafis, N.S.M. and Awang, S. (2021) 'An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification', *IEEE Access*, Vol. 9, pp.52177–52192.

Oyewole, G.J. and Thopil, G.A. (2023) 'Data clustering: application and trends', *Artificial Intelligence Review*, Vol. 56, No. 7, pp.6439–6475.

Qaiser, S. and Ali, R. (2018) 'Text mining: use of TF-IDF to examine the relevance of words to documents', *International Journal of Computer Applications*, Vol. 181, No. 1, pp.25–29.

Singh, S. and Mahmood, A. (2021) 'The NLP cookbook: modern recipes for transformer based deep learning architectures', *IEEE Access*, Vol. 9, pp.68675–68702.

Tian, Y., Li, W., Wang, S. and Gu, Z. (2023) 'Semantic similarity measure of natural language text through machine learning and a keyword-aware cross-encoder-ranking summarizer – a case study using UCGIS GIS &T body of knowledge', *Transactions in GIS*, Vol. 27, No. 4, pp.1068–1089.

Wang, H. (2025) 'The value orientation clustering analysis based on topic models in the social network environment', *International Journal of Information and Communication Technology*, Vol. 26, No. 8, pp.19–34.

Xing, L., Vadrevu, V.S.P.K. and Aref, W.G. (2025) 'The ubiquitous skiplist: a survey of what cannot be skipped about the skiplist and its applications in data systems', *ACM Computing Surveys*, Vol. 57, No. 11, pp.1–37.

Yang, J., Liu, Z., Xiao, S., Li, C., Lian, D., Agrawal, S., Singh, A., Sun, G. and Xie, X. (2021) 'Graphformers: GNN-nested transformers for representation learning on textual graph', *Advances in Neural Information Processing Systems*, Vol. 34, pp.28798–28810.

Yuan, Z., Chen, H., Zhang, P., Wan, J. and Li, T. (2021) 'A novel unsupervised approach to heterogeneous feature selection based on fuzzy mutual information', *IEEE Transactions on Fuzzy Systems*, Vol. 30, No. 9, pp.3395–3409.

Zheng, Z., Huang, S., Weng, R., Dai, X-Y. and Chen, J. (2020) 'Improving self-attention networks with sequential relations', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp.1707–1716.