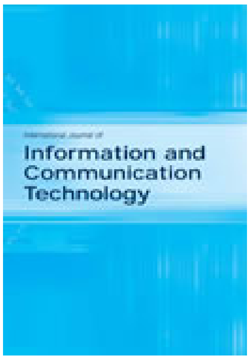


International Journal of Information and Communication Technology

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

Student career guidance using sentiment analysis and decision tree models

Yang Liu, Hengxing Kou

DOI: 10.1504/IJICT.2025.10073893

Article History:

Received:	30 June 2025
Last revised:	23 July 2025
Accepted:	23 July 2025
Published online:	22 October 2025

Student career guidance using sentiment analysis and decision tree models

Yang Liu and Hengxing Kou*

School of Innovation and Entrepreneurship,
Nantong Institute of Technology,
Nantong 226001, China
Email: m13855308244@163.com
Email: m18921468310@163.com
*Corresponding author

Abstract: Accurately analysing college students' career trends and offering individualised advice has become a significant difficulty to raise the quality of employment as the number of graduates keeps increasing and the employment scenario gets more and more complicated. Conventional career advice largely depends on one data source, so it produces inadequate career planning. This paper suggests SADTM-CIAG, a model for analysing and guiding college students' career inclination based on the combination of sentiment analysis and decision tree (DT), which is combined with personalised counselling strategies to give students scientific and reasonable career development advice, in order to solve this problem. This work designs two technical performance verification experiments and one application effect verification experiment to evaluate the efficacy of SADTM-CIAG in real-world career advising. According to the experimental data, the model has great application value and shows notable advantages in many respects.

Keywords: sentiment analysis; decision tree; DT; prediction of career inclination; personalised guidance.

Reference to this paper should be made as follows: Liu, Y. and Kou, H. (2025) 'Student career guidance using sentiment analysis and decision tree models', *Int. J. Information and Communication Technology*, Vol. 26, No. 38, pp.1–21.

Biographical notes: Yang Liu received his Master degree from Anhui Normal University in 2008. He currently serves as the Director of the Innovation and Entrepreneurship Teaching and Research Office at the College of Innovation and Entrepreneurship, Nantong Institute of Technology. His research interests include career planning guidance and entrepreneurship education.

Hengxing Kou received his Bachelor's degree from Nantong Institute of Technology in 2021 and is currently pursuing a Master's degree in Computer Science at Kanchanaburi Rajabhat University, Thailand. His research interests include career planning guidance and entrepreneurship education.

1 Introduction

High-quality and diverse abilities are increasingly sought after in society under the constant drive of economic change and technological innovation in the twenty-first century (Djekić et al., 2023). Simultaneously, the environment in which college students choose their career has grown more complex than it ever was. The job market of today not only demands graduates to have strong professional skills but also considers their whole quality, degree of innovation, and psychological adaptation level. When confronted with career decisions upon graduation, some college students show uncertainty, anxiety, and even blind choice due to insufficient knowledge of professional choice, blurred self-knowledge, and asymmetric social information; this results in a series of issues including poor employment quality, poor job matching, and increasing rate of change of career.

Higher education's basic goal is to encourage human beings' all-around development; professional development is therefore the essential link in the path of personal development. An important question in education management and student development support is how early, scientifically based career tendency of college students should be identified and how they can be helped to establish a clear career awareness and reasonable career goals. Trying to help students through psychological exams, career tests, interest questionnaires, and other tools, more and more schools and universities have been investigating career education and tailored counselling in recent years. These approaches still provide numerous difficulties for practical implementation, though, including single dimension, one-sided outcomes, weak interpretability, low adaptability, etc., which makes it difficult to satisfy the rising demand for differentiated development (Uysal et al., 2024).

In this framework, how to mix new information technologies means to better the scientific and focused identification of college students' vocational propensity has become a direction worthy of in-depth research. In public opinion monitoring, consumer behaviour analysis, and public management, sentiment analysis has been extensively applied recently. It can show deeper elements like an individual's attitudes, values, and natural motivation in addition to reflecting his or her emotional response to a certain event or object (Poria et al., 2020). Applied to the study of college students' career development, it can be utilised to understand the inclination reasons underlying their professional choices more fully by analysing people's emotional attitudes and psychological traits.

As a standard categorisation and decision-making tool, DT model also offers great execution efficiency, clear interpretability, and easy structure. DT is extensively applied in numerous educational prediction and recommendation systems for activities like course selection, student performance classification, and behaviour prediction (Khan and Ghosh, 2021). Its logical framework not only satisfies the teacher's demand for result interpretation but also helps coupled analysis with other feature data to offer a trustworthy basis for later tailored career advice.

This work attempts to build a model using sentiment analysis and DT technique for analysing and directing career inclination of college students. Representative variables are obtained as input features by aggregating data from the college student population in terms of career awareness, professional interest, psychological state, etc.; subsequently, the DT model is applied to forecast the career classification of students and output corresponding career advice and counselling strategies by combining their emotional tendencies and behavioural characteristics. Apart from attaining stratified and typed administration of students in career advisory services, the approach offers more focused customised education decision assistance for colleges and universities.

The significance of this study is mainly reflected in the following aspects: first, at the theoretical level, it enriches the research perspective of education informatisation and student development support, and provides theoretical support for the introduction of affective and decision-making intelligence in career education in colleges and universities; second, at the practical level, it provides a methodological basis for colleges and universities to build scientific, systematic, and efficient career guidance mechanisms; and third, at the technological level, it provides a methodological basis for subsequent use in other education services by combining the affective analysis with the interpretability model, it provides an experimental basis for subsequent migration applications in other educational prediction tasks.

In general, investigating the methodological path of integrating sentiment analysis and DT models not only responds to the realistic demand for career guidance in colleges and universities, but also offers a feasible solution for promoting the development of educational intelligence and the construction of personalised service system in the context of the current increasing demand for career decision support for college students. Theoretical and practical investigation of this study surrounding the data analysis method, model development process, experimental outcomes and strategy recommendations will methodically evolve in the next chapters.

2 Sentiment analysis

In the junction of natural language processing, artificial intelligence, and mental computation, sentiment analysis which is a crucial area of research with primarily the objective of identifying and quantifying an object's emotional state or subjective attitude in a given context. Sentiment analysis has evolved from the early task of text polarity classification to a complex modelling problem covering multimodal inputs, multi-sentiment dimensions, and multi-scene adaptation, which has become an important technological support for understanding human behavioural patterns and simulating cognitive processes with the development of human-computer interaction technology and the demand for emotional intelligence.

Sentiment analysis typically consists, from a methodological perspective, in four main phases: data collecting, feature extraction, sentiment classification and sentiment quantification (Yue et al., 2019). Mostly from multimodal information sources like text corpus, speech signals, visual expressions, behavioural trajectories, etc., data acquisition extracts raw samples including emotional cues. Based on data types, including TF-IDF word frequency features, word embedding vectors, picture convolutional features, or spectral characteristics of speech, the feature extraction stage generates vectors that can characterise emotional aspects. While sentiment quantification further provides organised inputs for subsequent modelling in the form of numerical assignments, intensity scores, or probabilistic outputs, sentiment classification tasks typically involve classifying the emotional state of an object into predefined categories such as positive, negative, neutral, or more detailed categories such as pleasure, anxiety, disgust, surprise.

Two main groups define existing sentiment analysis techniques: rule-based systems and learning-based. Rule-based approaches derive emotional inclinations by annotating sentiment word polarity and intensity values and matching them with input data during analysis from manually created sentiment lexicons with semantic rules, including SentiWordNet, AFinN, NTUSD, and other lexicon systems (Jiao and Chen, 2022). When faced with semantic ambiguities, complicated changes in context, or non-linguistic data, these approaches have limited performance capabilities and cannot automatically adapt to new domain features; nonetheless, they have certain advantages in low-resource settings and good interpretability.

Learning-based methods, on the other hand, discover mapping associations between sentiment expressions and features from labeled samples, therefore making predictions with the aid of machine learning or deep learning models (Xu et al., 2020). While deep models such as convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory network (LSTM), and Transformer class models such as BERT, which emerged in recent years, can capture richer contextual semantic information, and have achieved significant performance improvement in sentiment classification tasks. Traditional machine learning methods such as Plain Bayes and support vector machines (SVM) usually depend on bag-of- words models or TF-IDF features for training. Particularly appropriate for unstructured text sentiment analysis tasks, these models provide significant adaptability and generalising capacity. Deep models still need to be balanced against performance and cost in actual applications, although they also have certain issues including significant reliance on large-scale annotated corpuses, high training cost, limited model interpretability, etc.

Sentiment analysis can also be accomplished in structured data environments including questionnaires, rating scales and other non-natural language forms of data. Participants assign scores on several markers of emotional dimensions, such as satisfaction, interest, identification, aversion, etc., and finally these scoring values can be combined into an overall emotional disposition score. Strong operability and explicit dimensions define this quantitative method, which is especially fit for application scenarios with unambiguous emotional labeling and defined assessment dimensions (Yannakakis et al., 2018). Under this kind of situation, the weighted average model helps one to express the emotional tendency score.

$$E_i = \sum_{j=1}^{n} w_j \cdot s_{ij} \tag{1}$$

$$\sum\nolimits_{j=1}^{n} w_j = 1 \tag{2}$$

where E_i indicates the complete emotional score of the i^{th} person on an object or event; s_{ij} is its rating on the j^{th} emotional dimension; w_j is the weight coefficient of the dimension. Changing the weight parameters helps one to reflect the relevance of many characteristics in the comprehensive evaluation, so producing a modelling impact nearer the true perceptual experience.

Sentiment analysis has also developed recently in handling varied data from several sources. Apart from conventional text input, image recognition technology can detect emotional cues in facial expressions, speech processing technology can extract parameters including intonation, speed of speech and rhythm as the basis for emotional judgement, and behavioural trajectory analysis technology can deduce emotional states by means of user activity patterns including click paths and dwell time. To increase the robustness and accuracy of sentiment detection, the multimodal sentiment analysis model is exactly to combine the above several inputs and achieve joint modelling by feature splicing, attention mechanism or feature selection mechanism.

Sentiment analysis also makes more use unsupervised learning and transfer learning. Applications include sentiment word expansion and cluster discovery such as automatic word polarity categorisation using sentiment co-occurrence frequency which often call for unsupervised approaches (Sanagar and Gupta, 2020). Conversely, transfer learning reduces the need for labelled data by transferring linguistic or sentimental knowledge learnt in big corpora to the target task in new domain tasks with small sample sizes. These approaches show more promise in application situations with erratic domain distribution and limited resources.

As a fundamental tool for comprehending human subjective experience in artificial intelligence systems, sentiment analysis has evolved its methodological framework from single-modal, shallow semantics to multi-modal, deep semantic fusion overall. Its modelling capacity and interpretability are crucial whether in the domains of user profiling, personalised suggestion, education assessment, intelligent Q&A, social media opinion monitoring and so on (Eke et al., 2019). Sentiment analysis will reach deeper emotional understanding and intelligent response in a wider range of scenarios by means of constant optimisation of the model structure and improvement of the training mechanism, so offering strong support for the construction of human-driven information systems.

3 Decision tree

Mostly applied for classification and regression applications, DT is a class of supervised learning models grounded on tree topology. By progressively splitting the data into several subsets, DT creates a hierarchical tree model with leaf nodes matching the last prediction outcomes (Sarwar et al., 2025). Its main benefits are that the model structure is explicit, the decision path is apparent, and it is quite simple to explain and grasp, which makes DT extensively utilised in data mining, pattern recognition and intelligent decision-making systems.

The fundamental issue with the building of DT is selecting the optimum characteristics and division points to maximise the influence of every division. The optimal division should concentrate the samples of the same class as much as feasible, therefore rendering the split subset as pure as feasible. Information entropy is extensively applied to evaluate the uncertainty or purity of the dataset. The entropy of a dataset D with m categories and a probability of category i is defined as p_i, thus:

$$H(D) = -\sum_{i=1}^{m} p_i \log_2 p_i \tag{3}$$

Information entropy gauges in a data collection the degree of ambiguity or confusion. Higher values of entropy point to more homogeneous distribution of categories in the data collection and more information uncertainty (Mishra and Ayyub, 2019). Conversely, an entropy of zero denotes that every sample in the dataset fits exactly one another and the dataset is totally pure.

Following feature A for division, the dataset is split into numerous subsets D_1, D_2, …, D_v, and the weighted sum of the entropy of every subset result as the conditional entropy after division. The information gain of feature A is defined as the pre-division entropy's difference from the conditional entropy:

$$IG(A) = H(D) - \sum_{j=1}^{v} \frac{|D_j|}{|D|} H\left(D_j\right) \tag{4}$$

The information gain captures the improvement in the dataset's classification effect resulting from feature A. Calculating the information gain of all potential features, DT chooses the feature with the highest information gain for segmentation. Though it streamlines the search process, this greedy approach cannot ensure the global optimal division.

The recursive division of nodes will keep on during the building of DT till the termination criteria are satisfied. Typical termination criteria are: the tree reaches a given maximum depth; the number of samples is less than a predefined threshold; or no effective division can be established; the samples in the node all belong to the same class. Called a full tree, the tree created in this manner usually performs well on training data but may be overfitted with noise and outliers, therefore compromising generalisation.

Post-pruning was devised as a means of preventing overfitting (Al-Behadili et al., 2020). First builds the complete tree following post-pruning, then simplifies the model by eliminating underperforming subtrees depending on cross-valuation or validation set results. Apart from improving the generalisation performance of the model, post-pruning also simplifies interpretability and lessens the decision route complexity.

Apart from information gain, the DT algorithm solves the issue of information gain preference for multi-valued features by including metrics such as information gain rate and Gini index. The C4.5 method optimises the partition by means of the information gain rate, which normalises the information gain and conversely, the Gini index is appropriate for binary segmentation and employed in the CART algorithm to evaluate purity by computing the sum of squared category probability (Widowati, 2024).

DT has several benefits: first, it can directly deal with categorical and continuous features; second, the decision path is clear, which makes it easy to explain the model results; third, the training and prediction computations are rather small, which makes it appropriate for large-scale data processing. It does not need complicated normalisation or standardisation of data.

DT has flaws, though. The greedy splitting method depends on the local optimum and is readily influenced by noise, which shows up as overfitting. Particularly in situations with several complicated features, a single DT could not be sufficient. Furthermore, susceptible to the class imbalance issue are DTs, which can readily cause prejudice towards the majority class.

Integrated approaches which include random forest (RF), and gradient boosted tree (GBDT) are extensively used to get beyond the restrictions of a single DT. By training several DTs with random sampling and random feature selection, utilising a voting mechanism, RF increases stability and accuracy. Conversely, GBDT iteratively trains the DTs in a sequential fashion, with each tree fitting the residuals of the one before it, hence progressively optimising the total model error and proving great prediction capability (Kiran et al., 2023). These approaches considerably extend the DT applicability range.

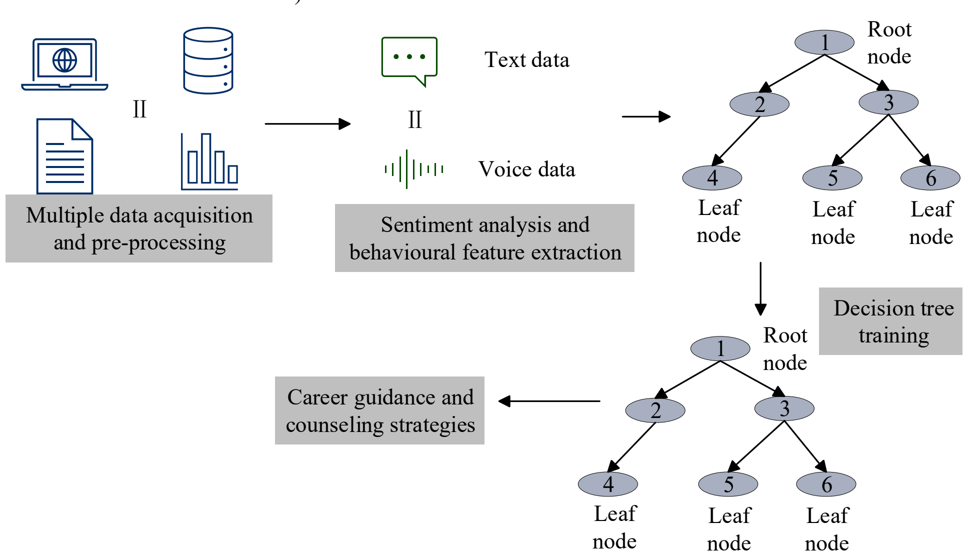
Although DT algorithms have a lengthy history of research, the traditional algorithms consist in ID3, C4.5, and CART. Based on information gain, ID3 is appropriate for discrete features; C4.5 introduces the information gain rate and enables continuous features and pruning; CART supports the Gini index and builds binary trees, which is ideal for classification and regression tasks. Improvement of model performance depends on selecting suitable algorithms and parameters.

With its straightforward, easy-to-understand and versatile characteristics, DT generally plays a significant role in the field of data analysis and machine learning. DT can function effectively in many complicated datasettings when combined with trimming and integration methods.

4 A model for analysing and guiding university students' vocational orientation

This work aims to build a model merging sentiment analysis and DT algorithms, called SADTM-CIAG, to analyse and guide college students' career inclination. See Figure 1 to allow colleges and universities to get accurate predictions of students' career inclination and personalised career guidance by multi-dimensional data analysis. There are four primary divisions to the fundamental process of the model:

Figure 1 Model for analysing and guiding university students' career inclinations (see online version for colours)



First, high-quality input for further research are obtained by gathering and cleansing the career cognition, professional interest, psychological condition, and other data of the college student group. Second, sentiment analysis and behavioural feature extraction extract students' emotional and behavioural tendencies, which together with other factors influencing occupational proclivities comprise the occupational categorisation input feature set. Third, utilising the DT algorithm, the DT model training and occupational trend classification prediction component classifies and projects student data and occupational trends. In the section on career guidance and counselling strategy output, the prediction findings of the model are applied to provide customised advice and counselling strategies to assist in career planning for individuals.

4.1 Data acquisition and pre-processing

It is a basic component of the SADTM-CIAG model, which seeks to offer high-quality input data for next sentiment evaluations and forecasts of career paths. Collecting students' behavioural data, affective data, career perceptions and other relevant information from many sources is part of this stage; then, cleaning, standardising, and feature extraction these data will guarantee data consistency and efficiency.

First, multidimensional data obtained via questionnaires, online learning environments, social media, and student behavioural records are gathered. The social media and online platform data comprise students' text and voice data, thereby expressing their affective inclinations and expressions towards career development. The questionnaire survey consists of questions and answers concerning students' job interests, professional perceptions, and career planning. To guarantee the generalisability and scalability of the research findings, the data collecting encompassed student groups with diverse disciplinary backgrounds, grade levels and geographic locations so ensuring the representateness of the sample data.

First, in pre-processing is data cleaning. Calculating the degree of deviation of the data points from the data means helps the Z-score standardisation method find and eliminate outliers greatly deviating from the general data distribution (Sullivan et al., 2021). The Z-score is computed by the formula:

$$Z = \frac{X - \mu}{\sigma} \tag{5}$$

where X is the data point's original value; μ is the mean value of that data point; σ is its standard deviation. A data point is considered an outlier and deleted if its Z value exceeds 3 or falls less than –3. This procedure guaranteed the correctness of the next analysis and practically removed noise from the data.

The k-nearest neighbour (KNN) technique was applied to fill in missing data. The KNN algorithm computes the similarity of the missing values with other sample points, therefore fills in the blank values using the values of the surrounding known data points. The KNN algorithm's fundamental concepts are to estimate missing values depending on known values of the nearest K neighbours selected based on the distance measure (Lee and Styczynski, 2018). KNN's formula is presented below:

$$\hat{y}_i = \frac{1}{K} \sum_{k=1}^{K} y_{i,k} \tag{6}$$

where K is the number of neighbours, $\hat{y}_i$ is the expected value of the i^{th} sample and $y_{i,k}$ is the known value of the K neighbours. Missing data points are filled in depending on the characteristics of their surrounding samples.

Apart from numerical characteristics, one-hot encoding is used for transformation in category-based data (graduates, majors, career impressions, etc.). One-hot encoding converts category variables into a set of binary features therefore enabling the model to grasp this category information apart from their original values.

Regarding feature extraction, behavioural data analysis helps to identify representative features among the pupils. Students' psychological condition and interest patterns might be reflected in their career awareness, professional interests, and participation in extracurricular activities. Effective hints about students' career preference can come from their participation in career planning events and frequency of talking about career development with colleagues.

Feature fusion forms the penultimate stage of data preparation. Integration of data from several sources results in a homogeneous dataset with multiple characteristics. Principal component analysis (PCA) technique was applied to downscale the features, so preserving the most instructive feature subsets and lowering the computational cost, so increasing the expressive capacity of the feature set.

The foregoing pre-processing and data collecting actions guarantee the quality and consistency of the input data, thereby laying a strong basis for the later sentiment analysis, behavioural feature extraction, and training of the occupational tendency prediction model. This procedure not only guarantees excellent data foundation for the correctness and dependability of the model but also enhances the availability of data.

4.2 Sentiment analysis and behavioural feature extraction

One of the main modules of the SADTM-CIAG model is emotional analysis and behavioural feature extraction, which seeks to extract useful characteristics from students' affective expressions and behavioural data to give great support for career inclination prediction. First analysis of pupils' emotional data comes in this step. Students' emotional displays in textual, voice, or behavioural data help to model their affective orientations. Students' comments on social media, responses to online surveys, course comments, and the contents of career-related emails all help one to deduce this information.

Sentiment analysis's main goal is to find in various contexts students' mood swings and emotional polarity. Valence aware dictionary and sentiment reasoner (VADER), a sentiment lexicon approach, was applied to measure the students' emotive inclinations by computing the sentiment scores in every text passage in the sentiment analysis of textual data process (Telmo et al., 2024). The sentiment score is computed with this formula:

$$S_{sentiment} = \sum_{i=1}^{n} \left(w_i \cdot t_i \right) \tag{7}$$

where $S_{sentiment}$ is the sentiment score; w_i is the sentiment weight of sentiment vocabulary t_i; n is the count of sentiment vocabulary. By weighing the sentiment words to mirror the emotional inclinations in a text, the method can determine its sentiment score. This helps one to pinpoint kids' emotional states in particular situations, including anxiousness, either positive or negative ones.

Sentiment analysis can be used on speech data of students in addition to textual data. One can further extract students' emotional traits by, for instance, analysing their intonation, speed of speech, and emotional intensity in self-reports or online interviews. Important foundations for behavioural feature extraction include also student behavioural data like career awareness, professional interests, and frequency of participation in different career-related events. One can get students' interest in inclinations and behavioural patterns by means of analysis of their social contacts, learning records, and involvement in career-related events.

Combining students' expression of interest, participation in career-related activities, and motivation in course learning, behavioural feature extraction uses the behavioural pattern mining technique to extract representative behavioural traits. These behavioural traits not only capture students' professional preferences but also expose their opinions and impressions on prospective career paths. For instance, a student demonstrates a strong interest and concern for job development if they regularly discuss workplace concerns on social media and show a great degree of participation in events connected to career preparation.

They must be combined into a single input feature set once the emotional and behavioural aspects have been isolated. All feature data must be normalised if we are to guarantee consistency and comparability across them. The min-max standardisation approach is applied to scale the feature values consistently to the [0, 1] interval during the standardisation phase (Mazziotta and Pareto, 2022). This is the standardisation formula:

$$X_{norm} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{8}$$

where X is the original data; X_{min} and X_{max} are respectively the minimum and maximum values of the feature. This approach guarantees consistent data scale and improves model stability and accuracy. All features have the same scale and value range after standardisation, therefore avoiding the scale variation between several features that influence the model training.

Following sentiment analysis and behavioural feature extraction, the dataset includes psychological state, career interest, and emotional inclination of the pupils. These characteristics will be inputs for the SADTM-CIAG model to assist the later prediction of career inclination and tailored advice policies. Combining affective analysis with behavioural data allows students' job preferences to be more fully evaluated, hence guiding the creation of individualised career advice programs.

4.3 Decision tree training and classification prediction of career preferences

At this point, the SADTM-CIAG model is trained on already fused feature data using the DT method. By using classification, the DT model forecasts the career inclination categories of students by using their affective traits, psychological condition, and professional interests, thereby offering them individualised career advice. This procedure mostly consists of several stages: DT training, feature classification, model prediction and outcome evaluation.

Analysing the multidimensional data features of students, the DT algorithm creates a tree-like structure in which every node denotes a segmentation condition of a feature, every branch corresponds to a different value of the feature, and every leaf node denotes a projected occupational category. By recursively segmenting the data and classifying samples in the same category to the same leaf node as much as feasible, the DT seeks to precisely forecast students' vocational preferences.

The SADTM-CIAG model selects the Gini index as the splitting criterion for DT in the training stage to guarantee that the model has low risk of overfitting when separating the features and good accuracy. The Gini index formula is:

$$Gini(D) = 1 - \sum_{i=1}^{c} p_i^2 \tag{9}$$

where p_i indicates the percentage of samples in the i^{th} class in the dataset D; c is the number of classes. DT chooses the features with the lowest Gini index for data division to increase the accuracy of classification since the purity of the dataset is higher in the smaller Gini index (Tangirala, 2020).

DT will choose the best features for division by recursion under the training process till the stopping criterion is met. Typical stopping criteria comprise the minimum number of leaf node samples, the maximum tree depth, etc. These constraints guarantee the generalisation capacity of the model by efficiently avoiding the overgrowth of DT and so lowering the occurrence of overfitting problems.

New student data can be utilised to forecast occupational inclination by means of the DT model following training. The new student data will start from the root node of the DT and judge along the branches of each node until it finally reaches the leaf node in the prediction stage; the occupational category corresponding with the leaf node is the prediction outcome of the student's occupational inclination.

The model uses the cost-complexity pruning approach to enhance the generalisation capacity of DT and prevent overfitting. By adding the pruning parameter α, this method regulates the complexity of the tree. Pruning aims to penalise the size of the tree so avoiding over-complexity of the DT and so enhancing its predictive capacity on fresh data. Pruning's cost complexity formula looks like this:

$$C_\alpha(T) = R(T) + \alpha|T| \tag{10}$$

where $C_\alpha(T)$ is the cost complexity of DT; $R(T)$ is the error rate (training error) of tree T; α is the penalty factor; $|T|$ is the number of nodes of tree T. Changing α helps one to manage the strength of pruning so balancing the appropriate ability and complexity of the model to obtain more accurate prediction.

By means of this training and prediction procedure, the SADTM-CIAG model can efficiently identify students' career trends, offer accurate career advice and development strategies for every student, enable colleges and universities to create more individualised career development guidance programs.

4.4 Career guidance and counselling strategy outputs

Following DT model training and career inclination classification prediction, the SADTM-CIAG model may offer individualised career advising and counselling plan output for every student. Based on the outcomes of the student's career inclination prediction, combined with the multi-dimensional analysis of his/her emotional characteristics, psychological state, professional interests, etc., the output of career guidance and counselling strategies generates customised career development suggestions to help the student identify the most appropriate career paths, formulate developmental goals, and provide corresponding support strategies.

First, the findings of the classification of students' career orientations will determine the output of career coaching. Every student's professional inclination falls into one of several categories (e.g., managerial, technical, creative, etc.). The system will automatically match the related job sectors and industry trends depending on the categories they fall into. These suggested job paths will fit students' professional interests, personality qualities, emotional condition, other elements, so offering them more scientific and logical career planning.

Second, the SADTM-CIAG model generates counselling techniques depending on the traits of students with distinct career orientations. For students who are inclined to technical careers, the model will advise relevant technical skills training, programming learning paths and industry dynamics to help them enhance their necessary professional competencies; for students who are inclined to management careers, the system will advise them to improve their leadership, teamwork and project management skills, and at the same time, provide them with relevant internships and career mentoring support. Every coaching approach is based on the individualised facts and professional choices of the student, meant to improve their career competitiveness.

The SADTM-CIAG model will not only offer workplace skills suggestions based on career inclination but also mix the findings of sentiment analysis to help students comprehend the influence of their emotional state on their career decisions in the strategy output process. The model will offer mental health adjustment recommendations to those students with high emotional swings to help them to evaluate their own needs and emotional tendencies more rationally in the process of career choice, so avoiding too strong influence of emotional elements on career decisions.

Furthermore, the SADTM-CIAG model can support career guidance departments of universities in making decisions by enabling them to adjust the administration of employment services depending on the data on the career orientations of the student population. To raise the employment rate and quality of employment of students, the system can assist schools in developing highly targeted employment training programs, optimising curriculum, planning internships fit for students' career interests, and offering real-time job market data.

The system is built considering several assessment methods to guarantee the efficacy of the output strategies. Tracking students' comments, job status, and progress in career development helps the model to constantly maximise the accuracy of career advice and counselling tactics and make dynamic changes when needed (Gati and Kulcsár, 2021).

Therefore, the SADTM-CIAG model additionally uses a career advice weighting algorithm to objectively ascertain the strategic weights of students' career direction. This algorithm helps the system automatically customise the best counselling technique for the individual by combining career inclination categories with the traits of the student in several angles. Here is the formula:

$$S_{guidance} = \sum_{i=1}^{n} w_i \cdot f_i \tag{11}$$

where $S_{guidance}$ is the weight of the final career advice; w_i is the weight of the i^{th} feature; f_i is the degree of influence of the i^{th} feature in the counselling strategy; n is the total weight of all the considered features.

The system generates more tailored career guidance by varying the weights and degree of influence of features thus ensuring that the student's career counselling strategy is exactly matched to his/her features, so boosting the possibility of her/his career development success.

By means of thorough analysis of their job choices and combination with tailored career development assistance and counselling tactics, the SADTM-CIAG model helps students make more scientific and logical decisions about their career routes. Simultaneously, it advances students' professional growth and social adaptation and offers more accurate career advice tools for schools and institutions.

5 Experimental design and results

5.1 Experimental dataset and assessment indicator setting

This study makes use of a self-constructed dataset, so data collecting took place between September 2024 and March 2025. Covering multi-dimensional aspects of students' career awareness, professional interests, psychological state and emotional expression, the data mostly consists of questionnaires, online learning platforms, social media, and student behaviour logs; the information is shown in Table 1. The variety and representativeness of the statistics are guaranteed by the data samples covering the college student population in several fields, grades, and geographical areas.

Table 1 Dataset information

Data source	*Data type*	*Main content*	*Feature dimensions*
Survey data	Student self-reported data	Includes career cognition, professional interests, emotional states, and psychological traits	Career cognition, professional interests, personality traits, learning motivation
Online learning platform	Learning behaviour data	Data on students' progress, participation, interactions, and emotional feedback	Learning progress, emotional tendencies, participation rate, interaction frequency
Social media data	Text and interaction data	Students' public statements and social interactions on social platforms	Emotional tendencies, interest changes, social behaviour, emotional fluctuations
Student activity logs	Behavioural log data	Logs of daily activities, class attendance, and learning time	Learning time, activity frequency, behaviour patterns, task completion

Every piece of data was carefully pre-processed to guarantee dependability and quality. About 5,000 college students' multidimensional information which includes over 5,000 valid data records spanning student populations of various disciplinary histories, grades, and geographical locations is gathered.

This study used three normalised evaluation indicators during the SADTM-CIAG model assessment process to guarantee a thorough and accurate assessment of model performance:

The normalised career prediction consistency (NCPC) is a metric of the consistency between the actual career choices of individuals and the anticipated categories of pupils given their preferences (Mahboob et al., 2024). With a bigger value, the match between prediction and reality is higher; this measure contrasts the degree of match between students' expected career categories and their final work areas.

Normalised sentiment regulation effectiveness (NSRE) measures its intervention effect on students's emotional fluctuation during the professional decision-making process (Han and Xu, 2024). Students' emotional comments during the counselling process help to evaluate this indicator; the degree of emotional swings before and after control is computed; larger values indicate more important impacts of emotional control.

Normalised guidance strategy adaptability (NGSA) evaluates how closely the model's produced personalised guidance plan fits students' actual career progress (Motwani et al., 2020). With higher scores indicating that the guidance strategy is more fit to students' particular needs, this indicator gauges the adaptability of the guiding strategy through students' comments on their actual career choices, employment situation, and career development trajectory after the guidance.

These three assessment measures give quantitative basis for later model refinement and application and can fairly represent the actual performance of the SADTM-CIAG model in terms of career inclination prediction, emotional regulation and counselling strategy generating.

5.2 Technical performance validation

The first experiment sought to confirm if the SADTM-CIAG model might benefit from the sentiment features acquired by the module of sentiment analysis. The increase of the predictive capacity of the models by the affective elements was evaluated by comparing the performance of the several models (see Table 2) on the prediction of occupational orientation.

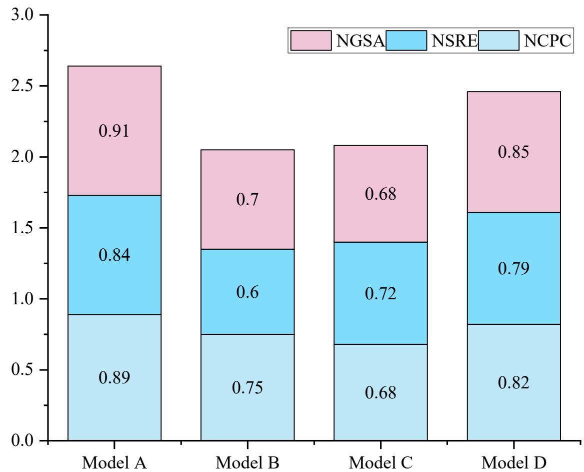
Table 2 Information of the four models

Model	*Description*
Model A (complete model)	A model that integrates sentimental features with student behavioural characteristics (such as career cognition, professional interest, learning motivation, etc.) for career tendency prediction. Sentiment features extracted by the sentiment analysis module are input together with other student behavioural features into DT model for training
Model B (control model 1)	A control model that does not use sentimental features but only uses traditional student features (such as learning behaviour, professional interest) as input to the DT model for prediction
Model C (control model 2)	A sentiment-based independent sentiment classification model, which uses only sentimental features to predict the student's career tendency, without considering other behavioural features
Model D (hybrid model)	A model that combines sentimental features with other student behavioural features. Different feature combinations (e.g., adding academic performance data or social behaviour data on top of sentiment features) are used for career tendency prediction

This experiment consists of the following: First the data is cleaned, including noise removal and handling missing values. Students' text data is subjected to sentiment analysis module to obtain sentiment features including sentiment scores and sentiment fluctuations. Standardised other behavioural traits guarantee comparability and consistency of data. Cross-validation raises model stability and helps to reduce overfitting issues. The same training and test sets were used for validation of every model.

The four models are next trained independently and assessed using cross-validation. Finally, three normalised evaluation criteria help to assess the performance of every model. The actual influence of affective elements in the prediction of vocational orientation is investigated by means of comparative performance of every model. Figure 2 shows the experimental outcomes.

Figure 2 Experimental results for experiment 1 (see online version for colours)

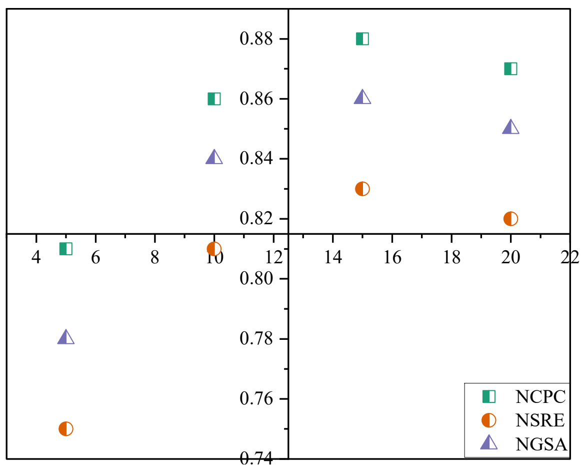


Based on the experimental data, Model A (SADTM-CIAG model) shows the best performance on all three evaluation criteria, particularly on NCPC and NGSA, which correspondingly reach 0.89 and 0.91 respectively. This implies that the SADTM-CIAG model, which combines emotional and other student behavioural characteristics, can more precisely forecast the vocational orientations of the students and offer more focused counselling techniques. Particularly in enabling the development of tailored counselling tactics, which is a major benefit over the conventional model, the influence of affective traits in the prediction of career inclination is well shown.

Models B and C did worse, on the other hand. Model B showed the favourable effect of affective qualities on career prediction since, although utilising just student behavioural traits, its NCPC of 0.75 was much lower than the entire model. Although its NSRE was higher (0.72), indicating that depending just on affective traits for occupational categorisation was not as effective as the full model, which combined other behavioural traits, Model C, which relied just on affective traits for prediction, had a lower overall NCPC (0.68).

These findings imply that the SADTM-CIAG model's effectiveness in career prediction and counselling tactics can be much improved by including emotional characteristics, therefore offering a strong theoretical basis for next studies. Experiment 2 looks at the effect of the maximum depth parameter of the DT on the career inclination prediction of college students to validate the technical performance of the SADTM-CIAG model even more. As a fundamental component of the structural complexity of DT, the maximum depth directly controls the degree of data detail the model can record and the feature level. We investigate its impact on the model generalisation ability and prediction accuracy by varying the maximum depth to find the best parameter setting to improve the effect of occupational inclination classification. The experimental stages are to establish varying maximum depth values (e.g., 5, 10, 15, 20) for DT model training on a preprocessed dataset including emotional and behavioural elements, accordingly. Using a cross-valuation approach, the predictive performance of the model is assessed for every depth configuration and the outcomes are expressed using uniform normalisation metrics (NCPC, NSRE, NGSA). By means of a comparison of the model performance at several depth values, the influence of the maximum depth on the prediction outcomes is investigated. Figure 3 shows the experimental outcomes.

Figure 3 Experimental results for experiment 2 (see online version for colours)



The model demonstrates notable improvement in the NCPC, NSRE, and NGSA measures as the maximum depth rises. As the maximum depth is raised from 5 to 15, SADTM-CIAG's performance keeps getting better since a deeper tree structure can better capture intricate elements in the data and thereby enhance prediction accuracy. But when the maximum depth is raised to 20, the performance somewhat declines which is probably because the model starts to overfit the training data, therefore affecting the generalisation capacity.

With a maximum depth of 15, the DT model obtained the optimum balance on the dataset of this study, therefore guaranteeing a good fitting capacity while avoiding overfitting. The outcomes offer a necessary base for later model optimisation and implementation. Combining multimodal feature fusion with this parameter setting would help to further enhance the predicted performance of the career inclination analysis and advice model for college students.

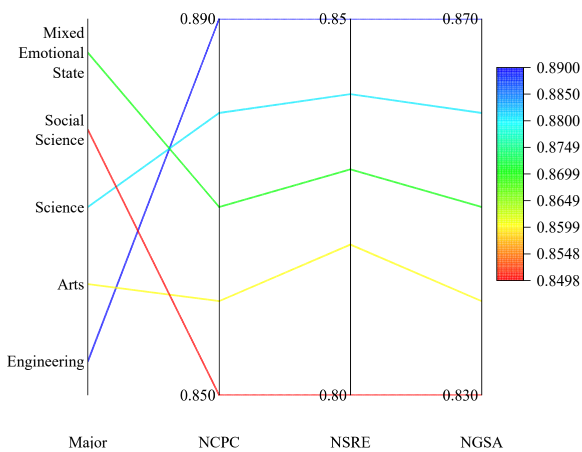
All things considered, these two studies confirm the stability and efficiency of the SADTM-CIAG model and its main parameter values in the analysis and direction of college students' career inclination, thereby providing a strong basis for the later construction of tailored career guiding methods. The next experiment will concentrate on the personalised adaptability of the output of the career advice strategies from the model to improve the practical value and application effect of the model even more.

5.3 Application effect verification

Experiment 3 seeks to confirm, in several student groups, the tailored adaptability and efficacy of the career guiding techniques produced by SADM-CIAG. To increase the relevance and efficacy of career advice, it is not only necessary to anticipate career inclination accurately but also to create customised counselling plans that fit students' emotional condition, interests, and psychological traits. By means of this experiment, we assess if the output of the guidance techniques from the model can satisfy the various demands of students with various backgrounds and emotional states.

Under consideration of the emotional fluctuations and student interests, a representative group of college students is chosen for the experiment covering engineering, science, liberal arts, social sciences, and other professional disciplines. Based on the career inclination projected by the SADTM-CIAG model, a matching tailored career guidance program is produced for every student group. The counselling effects are quantitatively analysed in the experiment using the normalised assessment indicators NCPC, NSRE and NGSA. Figure 4 illustrates the experimental outcomes.

Figure 4 Experimental results for experiment 3 (see online version for colours)



The SADTM-CIAG model provides more outstanding outcomes on all three measures, according to Experiment 3 findings for several groups of students with varying professional backgrounds and emotional states. On the NCPC indicator, engineering and scientific students scored correspondingly 0.89 and 0.89, respectively, suggesting that the model can better reflect their actual interests and potential and is more accurate in forecasting the career inclination of these two categories of students. The corresponding NSREs, at 0.85 and 0.84 respectively, are also higher, suggesting that the adjustment techniques meant for the emotional state of these students are successful and efficient and can help them better adjust their psychological state, so enabling the smooth progress of career planning.

Though slightly lower, at 0.86 and 0.85, the NCPC indicators for students majoring in arts and social sciences still show a high level, suggesting that the model is still more accurate in spotting the occupational inclination of these individuals. With 0.82 and 0.80 respectively, NSRE's validity is somewhat lower than that of the engineering and science groups; this group of students may have more complexity in emotional regulation, thus the model's designed regulating programs must be further polished to increase their efficacy. The NGSA indicators, all over 0.8, show that the course developed by the model has more practical value and adaptability and that diverse groups of students can accept and apply tailored counselling tactics better.

All the indicators in the group of mixed emotional states keep a high degree (NCPC is 0.87, NSRE is 0.83, and NGSA is 0.85), which totally indicates the good adaptation of the model to students with complicated emotional states. The outcome confirms the key function of sentiment analysis in the model, therefore enhancing the tailored matching and useful impact of the tutoring programme.

The SADTM-CIAG model shows excellent personalised adaptive capacity in emotional regulation and counselling implementation in addition to good accuracy in predicting career inclination. This approach offers colleges a sensible technical road to follow accurate and emotionally driven career advice. Later studies will keep improving the strategy design, increase the application possibilities of the model, and strengthen the whole influence of career advice.

6 Conclusions

6.1 Research summary and research shortcomings

This work systematically builds a multimodal data processing framework integrating emotional and behavioural traits, realises the classification prediction of students' vocational inclination based on the DT algorithm, and concentrates on the analysis and guiding strategy of college students' vocational inclination. By means of the thorough investigation of emotional state, professional interest, and psychological traits, the model successfully increases the accuracy of the prediction of career inclination and the relevance of personalised guidance, so offering strong technical support for vocational education and employment services in colleges and universities. The suggested SADTM-CIAG model shows improved performance in NCPC, NSRE, and NGSA, therefore confirming the practical worth and application possibilities of the model.

This study has some flaws even if the theoretical and practical levels show good outcomes. First, with restricted sample size and diversity, the data collecting mostly depends on questionnaires and online platforms, so influencing the generalising capability of the model. Second, the sentiment analysis module's emotion identification and behavioural feature extraction have not yet completely integrated more real-time dynamic data or adequately captured the intricate and shifting psychological condition of pupils. Furthermore, in high-dimensional and non-linear data the DT model still has significant restrictions even if it shows good predictive and explanatory ability.

Furthermore, the design of career guidance strategies is biassed towards the transformation of the expected results of the model, and the actual counselling pays insufficient attention to individual differences and counselling interactions to improve the flexibility and effectiveness of the strategies. Finally, the cross-campus and cross-regional applicability of the model has not yet been completely confirmed; also unclear are the generalisability of the model and its capacity to encourage its adoption.

6.2 Directions for future improvements

Future research can be improved in the following aspects:

1 Data diversity and scale expansion: real-time dynamic data (e.g., trajectory of learning behaviours, social interactions, etc.) should also be combined at the same time to reflect students' psychological and behavioural states more fully and to improve the generalising capacity and prediction accuracy of the model.

2 Model optimisation and multi-algorithm fusion: using DT, we investigate the effect of capturing non-linear correlations and the incorporation of sophisticated techniques such as integrated learning and deep learning to improve the processing capability of high-dimensional complex data.

3 Dynamic adjustment of personalised guidance strategies: strengthen the design of individual differentiation of career guidance methods, develop a feedback system, realise dynamic modification depending on students' feedback and behavioural changes, and improve the flexibility and adaptability of guidance techniques.

4 Cross-platform and cross-scene application promotion: promote the spread of the model to multi-scenario applications by investigating its applicability in many schools and institutions, diverse cultural backgrounds and career planning surroundings (Fu et al., 2025).

5 Multimodal sentiment analysis technology deepening: deepen the sentiment analysis technology, combine multimodal data including voice, image, text, etc., and enhance the capacity to detect and comprehend the complicated emotional state of pupils. Encourage the mix of cognitive psychology and affective computing to develop a more sophisticated emotional modelling system and improve the emotional sensitivity of the model and the degree of personalised service.

Although there are some flaws, overall, this study offers a creative technical path and practical paradigm for the analysis of college students' career inclination and personalised guidance. This lays a strong basis for the next in-depth investigation of the emotional driven career development support system.

Acknowledgements

This work is supported by the 2024 Jiangsu Province Higher Education Institutions Special Project on 'High-Quality Public Course Teaching Reform Research' (No. 2024GZJX016), the 2023 Provincial Higher Education Teaching Reform Project (No. 2023JSJG712), the 2023 Textbook Project Development: 'Career Planning and Employment Guidance for College Students' (No. 2023JJJ004) and the 2023 Nantong Institute of Technology Classroom Teaching Reform Project (No. 2023JKT059).

Declarations

All authors declare that they have no conflicts of interest.

References

Al-Behadili, H.N.K., Sagban, R. and Ku-Mahamud, K.R. (2020) 'Adaptive parameter control strategy for ant-miner classification algorithm', *Indonesian Journal of Electrical Engineering and Informatics* (*IJEEI*), Vol. 8, No. 1, pp.149–162.

Djekić, I., Velebit, B., Pavlić, B., Putnik, P., Šojić Merkulov, D., Bebek Markovinović, A. and Bursać Kovačević, D. (2023) 'Food quality 4.0: sustainable food manufacturing for the twenty-first century', *Food Engineering Reviews*, Vol. 15, No. 4, pp.577–608.

Eke, C.I., Norman, A.A., Shuib, L. and Nweke, H.F. (2019) 'A survey of user profiling: state-of-the-art, challenges, and solutions', *IEEE Access*, Vol. 7, pp.144907–144924.

Fu, T., Sun, R., Li, C. and Wang, L. (2025) 'Common phenomena exhibited by students in their individual design processes: a multi-scenario case study on software design education', *SAIEE Africa Research Journal*, Vol. 116, No. 2, pp.46–58.

Gati, I. and Kulcsár, V. (2021) 'Making better career decisions: from challenges to opportunities', *Journal of Vocational Behaviour*, Vol. 126, p.103545.

Han, Y. and Xu, Z. (2024) 'Fostering college students' mental well-being: the impact of social networking site utilization on emotion management and regulation', *BMC Psychology*, Vol. 12, No. 1, pp.1–15.

Jiao, J. and Chen, B. (2022) 'A GCN-and deep biaffine attention-based classification model for course review sentiment', *International Journal of Information Technologies and Systems Approach* (*IJITSA*), Vol. 16, No. 2, pp.1–18.

Khan, A. and Ghosh, S.K. (2021) 'Student performance analysis and prediction in classroom learning: a review of educational data mining studies', *Education and Information Technologies*, Vol. 26, No. 1, pp.205–240.

Kiran, S., Reddy, G.R., Girija, S., Venkatramulu, S. and Dorthi, K. (2023) 'A gradient boosted decision tree with binary spotted hyena optimizer for cardiovascular disease detection and classification', *Healthcare Analytics*, Vol. 3, p.100173.

Lee, J.Y. and Styczynski, M.P. (2018) 'NS-kNN: a modified k-nearest neighbors approach for imputing metabolomics data', *Metabolomics*, Vol. 14, pp.1–12.

Mahboob, K., Asif, R. and Haider, N.G. (2024) 'Career planning matters: Intelligence-based career path predictions using data mining models− a longitudinal study', *Mehran University Research Journal of Engineering and Technology*, Vol. 43, No. 4, pp.192–213.

Mazziotta, M. and Pareto, A. (2022) 'Normalization methods for spatio-temporal analysis of environmental performance: revisiting the min–max method', *Environmetrics*, Vol. 33, No. 5, p.e2730.

Mishra, S. and Ayyub, B.M. (2019) 'Shannon entropy for quantifying uncertainty and risk in economic disparity', *Risk Analysis*, Vol. 39, No. 10, pp.2160–2181.

Motwani, M., Soto, M., Brun, Y., Just, R. and Le Goues, C. (2020) 'Quality of automated program repair on real-world defects', *IEEE Transactions on Software Engineering*, Vol. 48, No. 2, pp.637–661.

Poria, S., Hazarika, D., Majumder, N. and Mihalcea, R. (2020) 'Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research', *IEEE Transactions on Affective Computing*, Vol. 14, No. 1, pp.108–132.

Sanagar, S. and Gupta, D. (2020) 'Unsupervised genre-based multidomain sentiment lexicon learning using corpus-generated polarity seed words', *IEEE Access*, Vol. 8, pp.118050–118071.

Sarwar, J., Khan, S.A., Azmat, M. and Khan, F. (2025) 'An application of hybrid bagging-boosting decision trees ensemble model for riverine flood susceptibility mapping and regional risk delineation', *Water Resources Management*, Vol. 39, No. 2, pp.547–577.

Sullivan, J.H., Warkentin, M. and Wallace, L. (2021) 'So many ways for assessing outliers: what really works and does it matter?', *Journal of Business Research*, Vol. 132, pp.530–543.

Tangirala, S. (2020) 'Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm', *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 2, pp.612–619.

Telmo, K.A.B., Alviola, K.V., Desamparado, J.J.S., Cabigan, J.N.A., Santiago Jr, C.S. and Shimada, R.A.A. (2024) 'Sentiment analysis of student evaluation for teachers using valence-aware dictionary and sentiment reasoner', *Journal of Interdisciplinary Perspectives*, Vol. 2, No. 9, p.1.

Uysal, A.S., Li, X. and Mulvey, J.M. (2024) 'End-to-end risk budgeting portfolio optimization with neural networks', *Annals of Operations Research*, Vol. 339, No. 1, pp.397–426.

Widowati, F.U. (2024) 'Application of C4. 5 algorithm with PSO feature selection and bagging technique on breast cancer classification', *International Journal of Management Science and Information Technology*, Vol. 4, No. 2, pp.312–320.

Xu, D., Tian, Z., Lai, R., Kong, X., Tan, Z. and Shi, W. (2020) 'Deep learning based emotion analysis of microblog texts', *Information Fusion*, Vol. 64, pp.1–11.

Yannakakis, G.N., Cowie, R. and Busso, C. (2018) 'The ordinal nature of emotions: An emerging approach', *IEEE Transactions on Affective Computing*, Vol. 12, No. 1, pp.16–35.

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