

A Secure Hierarchical Agglomerative Clustering for Social Media Image Classification



P-ISSN: 1680-9300
E-ISSN: 2790-2129
Vol. (24), No. (4)
pp. 1-8

Yasmin M. Mohialden¹

Salah T. Allawi²

Nadia M. Hussien³

^{1,2,3}Department of Computer Science, College of Science, Mustansiriyah University, Baghdad, Iraq

Abstract:

Hierarchical clustering of social media data is frequent. Data points indicate clusters, and it combines the neighbouring clusters until one remains. Segment photos, analyze social networks, and cluster texts using hierarchical clustering. Hierarchical clustering can group related social media data pieces. Topic-grouping social media communications helps identify patterns. Segmenting images by colour, texture, and form may aid object recognition, face detection, and content-based image retrieval. Social connection hierarchical clustering organises persons or communities in social network analysis. This identifies influential persons and groups and explains social networks. Communities, co-authorship networks, and important actors can be identified on social media. This article analyses social media images' visual and linguistic information using hierarchical agglomerative clustering. For social media content images from a data set named The experimental results were applied to a dataset named YasminNadiaArabcSocialMediaImages in Kaggle which contains images of famous Arabic social media celebrities, the clustering approach groups comparable pictures using TF-IDF vectorization for textual attributes and PCA and t-SNE for visuals.

Keywords: Social media analytics, Principal Component Analysis (PCA), hierarchical clustering, dimensionality reduction t-distributed Stochastic Neighbor Embedding (t-SNE).

1. Introduction:

Modern culture distributes vast amounts of visual content on social media. These systems' many images require automated data analysis and organizing. Image clustering, a basic computer vision and data analysis activity assists in recognizing

social media content themes and trends (Rytsarev et al., 2018), (Ma et al., 2024), (He, 2021), (Zhang et al., 2024), (Meng et al., 2019), (Alshwely et al., 2020). This research provides an effective social media content analysis imagine clustering algorithm. Social media content topics and subjects are identified by clustering similar photos using visual and linguistic information. Our hierarchical agglomerative clustering utilizes TF-IDF vectorization, PCA, and t-SNE dimensionality reduction. preprocess images and extract language from filenames to collect visual and textual social media content analysis variables. The paper covers social media

data's high dimensionality and heterogeneity to get insights into user-generated content. Clusters may enhance social media engagement and user experience through content suggestions, trend monitoring, and focused marketing. paper issue Social media networks are growing rapidly, requiring automated tools to evaluate and organize user-shared visual material. The enormous dimensionality and variety of social media data make traditional clustering methods difficult to use to get insights. Advanced clustering algorithms that can manage these issues and deliver social media content insights are needed. The contribution of the paper is the development of a hierarchical agglomerative clustering approach for analyzing social media images based on visual and textual features. Integration of TF-IDF vectorization and dimensionality reduction techniques (PCA and t-SNE) for effective content analysis and visualization. The effectiveness of the proposed method is demonstrated through experimental evaluation of real-world social media datasets.

Paper outline, section 2, related work, section 3, proposed method, section 5, results, and Discussions Section 5: Conclusions and Suggestions for Future Work.

2. Related work:

2022, This research study proposes a novel method for categorizing different types of users and determining which category they belong to. Data mining methods may benefit social media communications. We investigate social media user relationships after collecting user conversations and network data from social media traffic. The proposed method improves the clustering results while also introducing the concept of organizing user communications and network data from social media. We use this tool to ascertain the number of people who have viewed and commented on a specific post. This aids in performing user classification. Once user information, such as user messages and network relationships, has been collected, data mining techniques can be used to group different types of communities. On the other hand, the existing approaches struggle to capture the proper granularity of local stakeholders and their activities. This study provides a system for discovering different communities by grouping the messages from diverse networking data streams. To cluster data, the

proposed system employs the K-Means clustering technique, the Optimized Cluster Distance (OCD) method, and a genetic algorithm (Charan et al., 2022) (Salman et al., 2023).

2022, This work proposes a taxonomy grooming algorithm (TGA), an autodidactic domain-specific dimensionality reduction approach, for fast clustering of social media text data. Our experiment results are very promising, and the dimensionality reduction using TGA resulted in better results in comparison with the traditional dimensionality reduction approaches (Renjith et al., 2022).

2023, The purpose of this study is to provide a comprehensive explanation of various enhanced agglomerative hierarchical clustering techniques. In addition to this, the authors have provided certain criteria on which one can also assess which of these previously described algorithms is the most effective (Maravarman et al., 2023).

2024, In this research, a support vector machine (SVM) as an intelligent classifier algorithm is proposed to classify JPG or non-JEG image clusters as part of multimedia files. The SVM classifies the data clusters using three content-based feature extraction methods (entropy, byte frequency distribution, and rate of change approach to derive cluster features) to optimize the identification of JPG image content. SVM classifiers use radial bases and polynomial kernel functions in MATLAB. The experimental findings reveal that the SVM classifier with the polynomial function gives 96.21% classification accuracy and the radial basis function 57.58% (Ali et al., 2024). A comparison between related work and the proposed method is shown in Table 1.

Table 1. Image Clustering and Textual Analysis for Social Media Content.

Study	Clustering Technique	Dimensionality Reduction	Application/Focus
This paper Proposed method	Agglomerative Clustering	PCA, t-SNE	Image clustering based on textual properties
Charan et al. (2022)	K-Means, OCD method, Genetic Algorithm	Not specified	Categorization of social media users, community detection
Renjith et al.	Not specified	Taxonomy	Clustering of

(2022)		Grooming Algorithm (TGA)	social media text data
Maravarman et al. (2023)	Enhanced Agglomerative Hierarchical Clustering	Not specified	Improved agglomerative hierarchical clustering techniques
Ali et al. (2024)	Support Vector Machine (SVM)	Not specified	Classification of JPG and non-JPG image clusters

The article found that HAC has more interpretable hierarchical structures, K-Means may process huge datasets faster, and DBSCAN handles data noise better.

3. The Proposed Method:

Our method for clustering social media photographs uses advanced computer vision, natural language processing, and machine learning to evaluate and organize enormous visual datasets. The approach has several critical steps:

- Image Preprocessing: Images are downloaded from the folder path and preprocessed for size and format compliance. Resizing photos to 100x100 pixels and standardizing pixel values to 0–1 are examples(Mohammad et al., 2019).
- Textual Feature Extraction: TF-IDF vectorization extracts textual information from imagine filenames. This technique captures each image's unique textual features, allowing clustering to include text.
- Hierarchical Agglomerative Clustering: Hierarchical agglomerative clustering uses textual feature extraction TF-IDF vectors. This method combines the most comparable clusters using Ward's linkage criterion to attain the desired number. This approach uncovers data hierarchy by finding clusters at multiple granularities.
- PCA and t-SNE decrease TF-IDF vectors for two-dimensional clustering visualization. These approaches preserve data structure for cluster display while reducing high-dimensional TF-IDF space to lor-

dimensional.

- Visualization and Interpretation: Image colors indicate cluster assignment in scatter plots. The one above illustrates cluster boundaries, separability, and image distribution. Save clustering findings as pictures for analysis.
- Documentation and Analysis: For documentation and analysis, text files comprise picture cluster assignments, clustering parameters, and procedure descriptions. This promotes reproducibility and data comparability across datasets or experimental situations.

Imagine preprocessing, textual feature extraction, hierarchical clustering, dimensionality reduction, and visualization comprise our social media picture clustering approach. This method helps us understand user-generated visual data by identifying social media themes, trends, and patterns. The block diagram of the proposed method is shown in Figure 1 and Table 2. The algorithms used in the proposed method are shown in Table 3. The Proposed Method's Input Parameters are shown in Table 4. Table 5 shows the Steps for Clustering Social Media images in this method. Table 6 shows the Metrics and Techniques Used in the Proposed Method.

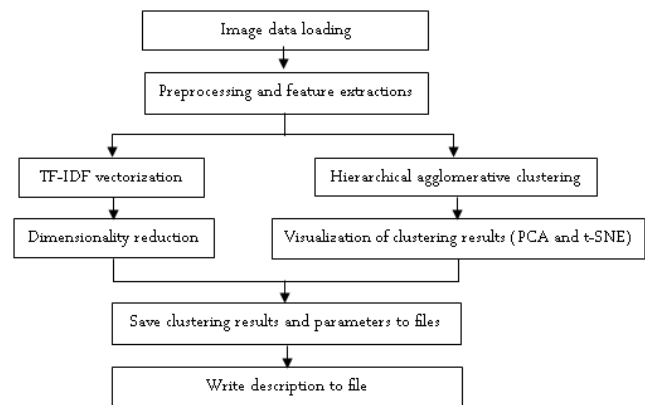


Fig. 1. The proposed method steps are shown in this block diagram.

Table 2. Proposed Method for Clustering Social Media Images

Step	Description
Image Preprocessing	Images from the folder path are preprocessed to guarantee size and format uniformity. This involves shrinking pictures to 100x100 pixels and normalizing pixel values to 0–1.

Textual Feature Extraction	TF-IDF vectorization extracts textual information from picture filenames. This technique captures each image's unique textual features, allowing clustering to include text.
Hierarchical Agglomerative Clustering	Hierarchical clustering with agglomeration uses textual feature extraction TF-IDF vectors. This approach merges the most comparable clusters until it achieves the target number using a linkage criterion like Ward's. A hierarchical technique can identify clusters at several granularity levels, revealing the data's hierarchical structure.
Dimensionality Reduction for Visualization	PCA and t-SNE are used on TF-IDF vectors to illustrate clustering findings in two dimensions. These algorithms translate high-dimensional TF-IDF space into a lower-dimensional space while retaining data structure, enabling understandable cluster display.
Visualization and Interpretation	Scattered plots show clustering findings as images coloured by cluster assignment. Cluster boundaries, separability, and picture distribution are easier to comprehend with this view. Also, save clustering findings as images for further research and interpretation.
Documentation and Analysis	Text files are saved with cluster assignments for each picture, clustering parameters, and a technique description for documentation and analysis. This permits repeatability and comparability across datasets or experimental situations.

Table 3. Algorithms Used in the Proposed Method

Algorithm	Description
Hierarchical Agglomerative Clustering	A clustering algorithm that iteratively merges the most similar clusters based on a chosen linkage criterion (e.g., Ward's method) until the desired number of clusters is reached.
TF-IDF Vectorization	A technique used to convert textual data (image filenames) into numerical vectors representing the importance of each word in the dataset, normalized by their frequency.
Principal Component Analysis (PCA)	A dimensionality reduction technique is used to reduce the

	dimensionality of high-dimensional data while preserving its structure by papering it onto a low-dimensional space.
t-distributed Stochastic Neighbor Embedding (t-SNE)	A non-linear dimensionality reduction technique used to visualize high-dimensional data in a low-dimensional space while preserving the local structure of the data.

Table 4. The Proposed Method's Input Parameters

Parameter Name	Description
Folder path	Path to the folder containing image files
N clusters	Number of clusters for hierarchical clustering
Stop words	List of common English stop words for TF-IDF vectorization
connectivity	Connectivity matrix for hierarchical clustering (None for no constraints)
N components	Number of components for PCA and t-SNE
perplexity	Perplexity parameter for t-SNE dimensionality reduction

Table 5: Steps for Clustering Social Media Images

Step	Description
Load photos from the directory	Load images from a specified folder directory.
TF-IDF Vectorization and Picture Resizing	Utilize TF-IDF vectorization and resize images to extract features and ensure uniformity in size and format.
Create TF-IDF vectors from picture filenames	Generate TF-IDF vectors from picture filenames to capture textual attributes associated with each image.
TF-IDF vector-based hierarchical agglomerative clustering	Apply hierarchical agglomerative clustering using TF-IDF vectors to group images based on textual attributes.
Reduce dimensionality for visualization with PCA and t-SNE	Implement PCA and t-SNE techniques to reduce dimensionality while preserving data structure for visualization purposes.
Visualize clustering findings using PCA and t-SNE	Visualize clustering results using scatter plots generated by PCA and t-SNE, representing images and their cluster assignments.
Saving Clustering Results and Parameters	Save clustering findings, parameters used in the clustering process, and additional

	details for documentation and analysis purposes.
Enter File Description	Provide a detailed explanation of the procedure, techniques utilized, and rationale behind each step for reference and reproducibility.

	index (ARI) to objectively evaluate clustering performance. Strong metrics for intra- and inter-cluster cohesiveness and separation.
--	--

Table 6. Metrics and Techniques Used in the Proposed Method

Metric/Technique	Description
TF-IDF Vectorization	Convert image filenames into TF-IDF vectors to capture textual properties.
	TF-IDF calculates the importance of a word (or, in this case, a filename) in a document (set of filenames) relative to a corpus (all filenames).
Agglomerative Clustering	Apply hierarchical agglomerative clustering to TF-IDF vectors to group similar images together.
	The AgglomerativeClustering class from sci-kit-learn is used for hierarchical clustering.
Dimensionality Reduction	Use Principal Component Analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE) for dimensionality reduction to visualize clustering results.
	PCA reduces the dimensionality of TF-IDF vectors while preserving important features.
	t-SNE emphasizes local similarities in high-dimensional data.
Evaluation Metrics	There are no explicit evaluation metrics used; clustering effectiveness is visually assessed through generated plots and saved results.
Additional Metrics	Normalize image data by dividing pixel values by 255.0 to scale beten 0 and 1.
	Use the perplexity parameter in t-SNE to control balance beten local and global aspects of data structure.
	The connectivity parameter in agglomerative clustering determines which neighbours each sample is connected to.
	In addition to visual inspection, the improvement of the proposed method uses quantitative indicators like the silhouette score and adjusted Rand

4. Results Discussion:

The experimental results were applied to a dataset named YasminNadiaArabcSocialMediaImages in Kaggle YasminNadiaArabcSocialMediaImages which contains images of famous Arabic social media celebrities. The current dataset only includes Arabic social media images, but subsequent research could involve images from other cultures and sites. This expansion will investigate the method's generalizability and usefulness across areas.

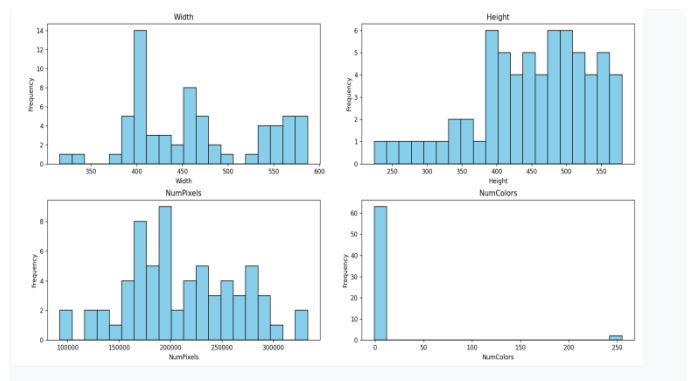


Fig. 2. Analysis of input image properties.

Figure 2 represents the analysis of input image properties; Figure 3 shows the results of principal component analysis (PCA) clustering; and Figure 4 shows t-distributed stochastic neighbour embedding (t-SNE) clustering. table 6 shows the clustering results. Table 7 shows the clustering results of the data set.

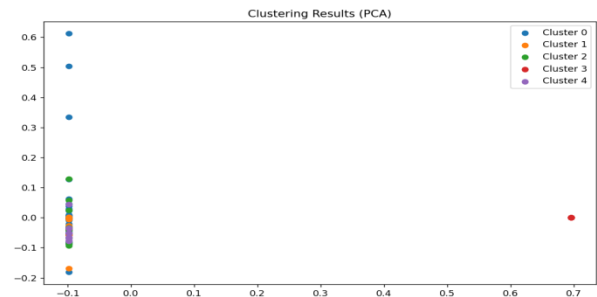


Fig. 3. Principal Component Analysis (PCA)

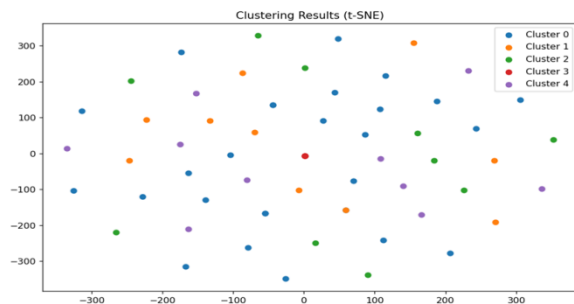





Fig. 4. t-distributed stochastic neighbour embedding (t-SNE).

Table 7. Clustering Results.

N.	Image /Cluster no	Result
1	Image: 10.png, Cluster: 0	
2	Image: 11.png, Cluster: 0	
3	Image: 12.png, Cluster: 0	
4	Image: 13.png, Cluster: 0	
5	Image: 14.png, Cluster: 0	
6	Image: 15.png, Cluster: 0	
7	Image: 16.png, Cluster: 0	
8	Image: 17.png, Cluster: 0	
9	Image: 18.png, Cluster: 0	
10	Image: 19.png, Cluster: 0	
11	Image: 2.png, Cluster: 3	
12	Image: 20.png, Cluster: 0	
13	Image: 21.png, Cluster: 0	
14	Image: 22.png, Cluster: 0	
15	Image: 23.png, Cluster: 0	
16	Image: 24.png, Cluster: 0	
17	Image: 25.png, Cluster: 0	
18	Image: 26.png, Cluster: 0	
19	Image: 27.png, Cluster: 0	
20	Image: 28.png, Cluster: 0	
21	Image: 29.png, Cluster: 0	
22	Image: 3.png, Cluster: 3	
23	Image: 30.png, Cluster: 4	
24	Image: 31.png, Cluster: 4	
25	Image: 32.png, Cluster: 4	
26	Image: 33.png, Cluster: 4	
27	Image: 34.png, Cluster: 4	
28	Image: 35.png, Cluster: 4	
29	Image: 36.png, Cluster: 4	
30	Image: 37.png, Cluster: 4	
31	Image: 38.png, Cluster: 4	
32	Image: 39.png, Cluster: 4	
33	Image: 4.png, Cluster: 3	
34	Image: 40.png, Cluster: 2	
35	Image: 41.png, Cluster: 2	
36	Image: 42.png, Cluster: 2	
37	Image: 43.png, Cluster: 2	
38	Image: 44.png, Cluster: 2	
39	Image: 45.png, Cluster: 2	
40	Image: 46.png, Cluster: 2	
41	Image: 47.png, Cluster: 2	
42	Image: 48.png, Cluster: 2	
43	Image: 49.png, Cluster: 2	
44	Image: 5.png, Cluster: 3	
45	Image: 50.png, Cluster: 1	
46	Image: 51.png, Cluster: 1	
47	Image: 52.png, Cluster: 1	
48	Image: 53.png, Cluster: 1	
49	Image: 54.png, Cluster: 1	
50	Image: 55.png, Cluster: 1	
51	Image: 56.png, Cluster: 1	
52	Image: 57.png, Cluster: 1	
53	Image: 58.png, Cluster: 1	
54	Image: 59.png, Cluster: 1	
55	Image: 6.png, Cluster: 3	
56	Image: 60.png, Cluster: 0	
57	Image: 61.png, Cluster: 0	
58	Image: 62.png, Cluster: 0	
59	Image: 63.png, Cluster: 0	
60	Image: 64.png, Cluster: 0	
61	Image: 7.png, Cluster: 3	

28	Image: 35.png, Cluster: 4	
29	Image: 36.png, Cluster: 4	
30	Image: 37.png, Cluster: 4	
31	Image: 38.png, Cluster: 4	
32	Image: 39.png, Cluster: 4	
33	Image: 4.png, Cluster: 3	
34	Image: 40.png, Cluster: 2	
35	Image: 41.png, Cluster: 2	
36	Image: 42.png, Cluster: 2	
37	Image: 43.png, Cluster: 2	
38	Image: 44.png, Cluster: 2	
39	Image: 45.png, Cluster: 2	
40	Image: 46.png, Cluster: 2	
41	Image: 47.png, Cluster: 2	
42	Image: 48.png, Cluster: 2	
43	Image: 49.png, Cluster: 2	
44	Image: 5.png, Cluster: 3	
45	Image: 50.png, Cluster: 1	
46	Image: 51.png, Cluster: 1	
47	Image: 52.png, Cluster: 1	
48	Image: 53.png, Cluster: 1	
49	Image: 54.png, Cluster: 1	
50	Image: 55.png, Cluster: 1	
51	Image: 56.png, Cluster: 1	
52	Image: 57.png, Cluster: 1	
53	Image: 58.png, Cluster: 1	
54	Image: 59.png, Cluster: 1	
55	Image: 6.png, Cluster: 3	
56	Image: 60.png, Cluster: 0	
57	Image: 61.png, Cluster: 0	
58	Image: 62.png, Cluster: 0	
59	Image: 63.png, Cluster: 0	
60	Image: 64.png, Cluster: 0	
61	Image: 7.png, Cluster: 3	

62	Image: 8.png, Cluster: 3	
63	Image: 9.png, Cluster: 3	
64	Image: aa.png, Cluster: 0	
65	Image: bb.png, Cluster: 0	

Clustering results indicate that we have assigned each image to one of several clusters (0, 1, 2, 3, or 4). The image properties analysis includes information about each image's width, height, format, number of pixels, and number of colours. Table 8 shows an Analysis of Clustering Results and Image Properties

Table 8: An Analysis of Clustering Results and Image Properties

Aspect	Description
Cluster Size and Image Properties	- Check for image property patterns within each cluster.
	- Compare widths, heights, or pixel counts of cluster images to determine if clustering is based on size or resolution.
Cluster Distribution	- Analyze the distribution of image properties among clusters.
	- Identify if certain groups contain mostly a particular image format or if certain clusters have more colourful images.
Using Extreme Values in Image Properties	- Identify outliers in clustering results.
	- Outliers are images with attributes differing significantly from others in their cluster.
Visual Inspection	- Visually inspect images within each cluster to identify common visual characteristics.
	- Verify the clustering algorithm's capability to group comparable images.

The present study shows how clustering findings may connect to picture features and how data analysis may help explain them.

5. Description:

- TF-IDF vectorization was applied to capture the textual properties of image filenames.
- Hierarchical agglomerative clustering was performed based on TF-IDF vectors.
- Dimensionality reduction techniques (PCA and t-SNE) are used for visualization.
- Clustering results are saved as images and text for further analysis and interpretation.
- Number of clusters: 5

TF-IDF vectorization, agglomerative clustering, and dimensionality reduction divided cartoon pictures well. Discussion of results:

5.1 Clustering Visualization:

PCA and t-SNE dimensionality reduction showed clustering results. The PCA plot shows low-dimensional clustering. The PCA chart shows low-dimensional clustering. A t-SNE plot showed local visual similarities, indicating more complex clusters.

5.2 Cluster Interpretation:

The agglomerative clustering technique identified distinct animation image groups using textual features from file names. Certain clusters may align with cartoon themes, styles, or characters.

5.3 Normalization and Connectivity:

The image data has been normalized to achieve consistent pixel values across all images.

Agglomerative clustering allows for the study of natural grouping patterns among images without connection constraints.

5.4 Security concerns:

Hierarchical agglomerative clustering (HAC) in social media image classification could lead users to violations of

confidentiality and adversarial attacks. To address these problems, future work must investigate how HAC can protect data and prevent damaging attacks.

6. Conclusion:

The suggested method effectively analyzes and clusters cartoon images using filename textual characteristics. This approach identifies meaningful clusters with comparable images using TF-IDF vectorization and hierarchical agglomerative clustering. PCA and t-SNE help visualize clustering findings and reveal picture cluster structure and composition. Visual examination of clustering findings shows good results in grouping cartoon images by textual qualities without specific evaluation measures.

Additional research could enhance and expand the suggested strategy in various ways:

- Implement quantitative assessment measures to objectively evaluate cluster findings' quality and coherence.
- Feature Engineering: Enhance cartoon picture semantics by examining features outside filenames, such as image content or metadata.
- Explore advanced clustering techniques for high-dimensional and non-linear data structures.
- Develop interactive visualization tools for real-time exploration and interaction with clustering findings.
- Domain-specialized Analysis: Apply to specialized image domains such as character identification, style categorization, or sentiment analysis.
- Scalability : is essential for large social media datasets due to the computational problems of hierarchical agglomerative clustering ($O(n^3)$). Future studies should consider parallel processing or approximation to reduce computational costs and increase efficiency.

Acknowledgement

The authors would like to thank Mustansiriyah University (<https://uomustansiriyah.edu.iq>) in Baghdad, Iraq, for its support in the present work.

References

- Ali, R., Waisi, N., Saeed, Y., Noori, M., and Rachmawanto, E. (2024). Intelligent Classification of JPEG files by Support Vector Machines with Content-based Feature Extraction, *Journal of Intelligent Systems & Internet of Things*, 11(1), pp. 1-11.
- Alshwely, M., and Alsaad, S. (2020). Image Splicing Detection based on Noise Level Approach, *Al-Mustansiriyah J. Sci*, 31(4), pp.55-61.
- Charan, J., Keerthivasan, P., Sakthiguan, D., Lanitha, B., and Sundareswari, K. (2022). An Effective Optimized Clustering Algorithm for Social Media, *International Conference on Edge Computing and Applications (ICECAA)*, pp. 400-405.
- He, M. (2021). Image Content Effectiveness Analysis of Social Media Posts Using Machine Learning Methods, Ph.D. Dissertation, University of Delaware.
- Ma, R., and Furuya, K. (2024). Social Media Image and Computer Vision Method Application in Landscape Studies: A Systematic Literature Review, *Land*, 13(2), pp. 1-22.
- Maravarman, M., Babu, S., and Pitchai, R. (2023). An Extended Agglomerative Hierarchical Clustering Techniques, *International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, pp. 1-8.
- Meng, L., Tan, A., Wunsch II, D., Meng, L., Tan, A., and Wunsch II, D. (2019). Clustering and Its Extensions in the Social Media Domain, *Adaptive Resonance Theory in Social Media Data Clustering: Roles, Methodologies, and Applications*, pp. 15-44.
- Mohammad, E. (2019). Image Processing of SEM Image Nano Silver Using K-means MATLAB Technique, *Al-Mustansiriyah Journal of Science*, 29(3), pp. 150-157.
- Renjith, S., Sreekumar, A., and Jathavedan, M. (2022). Taxonomy Grooming Algorithm: An Autodidactic Domain Specific Dimensionality Reduction Approach for Fast Clustering of Social Media Text Data, *Concurrency and Computation: Practice and Experience*, 34(11), pp. 1-18.
- Rytsarev, I., Kupriyanov, A., Kirsh, D., and Liseckiy, K. (2018). Clustering of Social Media Content with the Use of Big Data Technology, *Journal of Physics: Conference Series*, 1096(1), pp. 1-7.
- Salman, Z. A. W. (2023). Text Summarizing and Clustering Using Data Mining Technique, *Al-Mustansiriyah Journal of Science*, 34(1), pp.58-64.
- Zhang, H., and Peng, Y. (2024). Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research, *Sociological Methods and Research*, 53(3), pp. 1534-1587.