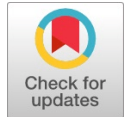


COVID-19 Sentiment Analysis using K-Means and DBSCAN

Smitesh D. Patravali, Siddu P. Algur



Abstract: The analysis of sentiment towards COVID-19 plays a crucial role in understanding public opinion. This research paper proposes sentiment analysis using K-means and DBSCAN clustering algorithms on the dataset of tweets related to COVID-19. Pre-processing and extraction of features is carried out using Term Frequency-Inverse Document Frequency (Tf-idf) to capture the weight of words in the dataset. K-means clustering is explored to group similar sentiments together, enabling the identification of sentiment clusters related to COVID-19. The DBSCAN algorithm is then employed to identify outliers and noise in the sentiment clusters. The evaluation metrics considered were accuracy, recall, F1-score, and precision. It was observed that DBSCAN was more effective in identifying underlying patterns in the data more accurately.

Keywords: COVID-19, K-Means, DBSCAN, social media, Public Opinion.

I. INTRODUCTION

The COVID-19 pandemic has significantly impacted societies worldwide, causing emotional and psychological effects on individuals. As people increasingly turn to social media platforms to express their thoughts, interpreting the sentiment of COVID-19-related content becomes crucial for public health trends. By applying sentiment analysis techniques to COVID-19-related content, we can gain insights into public sentiment, ranging from positive and hopeful sentiments to negative ones. The proposed work considers two popular clustering algorithms: K-means and DBSCAN. These unsupervised algorithms group similar sentiments together and identify distinct sentiment clusters within the data. The K-means partitions data into clusters based on similarity measures. It assigns data iteratively to the nearest cluster centroid and updates the centroid until convergence is achieved. DBSCAN categorizes data points based on their density within a given neighborhood. DBSCAN does not require the cluster count to be predefined. By applying K-means and DBSCAN clustering algorithms to COVID-19 sentiment analysis, we aim to achieve several objectives. First, to identify distinct sentiment clusters, such as positive, negative, neutral, or mixed sentiments, within the COVID-19 textual data. Second, we aim to analyze the

distribution and prevalence of sentiment clusters over time, enabling us to identify temporal trends in public sentiment. Furthermore, this study aims to compare the effectiveness of K-means and DBSCAN algorithms in clustering COVID-19 tweets. The remaining part of the paper is organized as follows: related work is discussed in Section 2. Section 3 provides an outline of the proposed work. In Section 4, the results of the work are discussed, and Section 5 concludes the paper with scope for the future work.

II. RELATED WORK

Smith et al. (2022) in [1] applied K-means clustering to analyze public sentiment towards COVID-19. The study provided insights into the different sentiment patterns and identified key topics of discussion. Johnson et al. (2022) in [2][19][20][21] investigated on clustering COVID-19-related tweets using K-means to perform sentiment analysis. They applied preprocessing and feature extraction methods to represent tweets as numerical vectors. Chen et al. (2021) in [3] aimed to capture the temporal issues of COVID-19 sentiment using K-means clustering. By applying K-means clustering, they identified distinct sentiment clusters over time, enabling a deeper understanding of sentiment trends and shifts during different phases of the pandemic. Lee et al. (2021) in [4] uncovered sentiment patterns in COVID-19-related social media posts. The findings revealed distinct sentiment patterns, helping to identify the key concerns, emotions, and opinions expressed by the public during the pandemic.

Wang et al. (2020) in [5] explored K-means clustering to analyze sentiment in COVID-19 related discussions on online forums. The study collected a large dataset of forum posts and applied sentiment analysis techniques to classify the sentiments expressed in the text. Gupta et al. (2020) in [6] employed K-means clustering to analyze individual perception of COVID-19 using Twitter data. The study provided insights into the different perspectives and emotions expressed by Twitter users regarding COVID-19. Patel et al. (2020) in [7] investigated on understanding COVID-19 sentiment using K-means clustering and social media data. The findings revealed distinct sentiment clusters, shedding light on the public's emotional response to the pandemic. Zhang et al. (2020) in [8] analyzed public opinion on COVID-19 by applying K-means clustering to data collected from online forums. Li et al. (2020) in [9] investigated mining COVID-19 sentiments using K-means clustering and news articles.

Manuscript received on 07 August 2023 | Revised Manuscript received on 10 October 2023 | Manuscript Accepted on 15 November 2023 | Manuscript published on 30 November 2023.

*Correspondence Author(s)

Smitiesh D. Patravali*, Research Scholar, Vijayanagara Sri Krishnadevaraya University, Ballari, Karnataka, India. E-mail: smitiesh.patravali@gmail.com

Dr. Siddu P. Algur, Ex. Vice Chancellor, Vijayanagara Sri Krishnadevaraya University, Ballari, Karnataka, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

Zhang et al. (2022) in [10][22][23] applied DBSCAN clustering algorithm to analyze sentiment in COVID-19-related social media data. Liu et al. (2022) in [11] investigated sentiment of COVID-19 using DBSCAN clustering. The study collected a large dataset of tweets and performed sentiment analysis on the text. Wang et al. (2021) in [12] employed DBSCAN clustering for unsupervised sentiment analysis of COVID-19 news articles. The researchers applied text processing techniques to extract sentiment features and used DBSCAN to group articles with similar sentiments. The findings provided insights into sentiment patterns in COVID-19 news coverage. Chen et al. (2021) in [13] explored DBSCAN for sentiment analysis of COVID-19 tweets.

Li et al. (2020) in [14] investigated sentiment analysis of COVID-19 discussions in online forums using DBSCAN clustering. The findings provided insights and diverse opinions expressed in online discussions. Zhao et al. (2020) in [15] employed DBSCAN clustering for temporal sentiment analysis of COVID-19 news articles. Liu et al. (2020) in [16] investigated sentiment clustering of COVID-19 social media data using DBSCAN. The researchers collected a large dataset of social media posts and performed sentiment analysis. Wang et al. (2020) in [17] explored DBSCAN clustering for sentiment of COVID-19 tweets. Liu et al. (2020) in [18] investigated sentiment analysis of COVID-19 news headlines using DBSCAN clustering.

III. PROPOSED SYSTEM DESIGN

The system design begins with data collection of COVID-19 related tweets. The collected tweets were subjected to pre-processing, tokenization, and lemmatization to ensure better analysis. The system design considers Tf-idf to represent the pre-processed tweets into numerical representation. By applying K-means to sentiment analysis, the system groups similar sentiments together, allowing for the identification of positive, negative, and neutral sentiment clusters. The DBSCAN algorithm offers the advantage of density-based clustering. The K-means algorithm provides an initial clustering structure, while DBSCAN enhances it by capturing additional sentiment nuances and identifying outliers. To evaluate the performance of the model, various metrics are calculated.

The proposed model has following components -

- Data collection
- Data pre-processing
- Unsupervised clustering model
- Clustering analysis

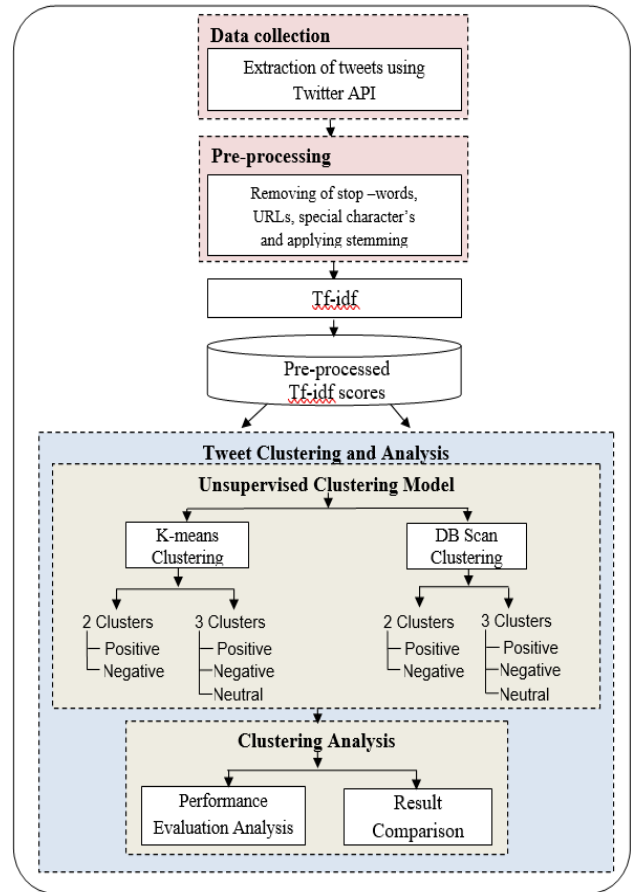


Figure 1. System Design for COVID-19 Analysis of Sentiments Using Unsupervised Learning

3.1 Data collection: The tweets related to COVID-19 pandemic were collected using the Twitter API from March 2020 to May 2022. The extraction of tweets was conducted using the Twitter API, which gathers real-time information shared on Twitter. The Twitter API offers a wide range of features that enable users to access and analyze tweets based on different criteria. Around 8000 tweets were extracted using API.

3.2 Data preprocessing: Preprocessing techniques are employed to clean and prepare the data for analysis. Removing stop words, which have minimal semantic meaning, helps reduce noise in the data. This typically involves eliminating unnecessary characters like punctuation marks and special symbols while ensuring consistent formatting throughout the text. Additionally, tokenization is performed to break the text into tokens. Stemming or lemmatization techniques is used to normalize words and bring them to their base forms. Data preprocessing involves handling negation, where the presence of negation words can invert the sentiment expressed in the text.

E.g. Consider the following tweet: 'Can't believe it's been a year since the first lockdown! 🤔 Time flies! #COVID19 #pandemic #lockdownlife'

Step 1. Initially, the tweet is tokenized as: ['Can't', 'believe', 'it's', 'been', 'a', 'year', 'since', 'the', 'first', 'lockdown', '!', '🤔', 'Time', 'flies', '!', '#COVID19', '#pandemic', '#lockdownlife']

Step 2. Further the exclamation marks, the mask emoji, and the hashtag symbols would be removed, resulting in: ['Can't', 'believe', 'it's', 'been', 'a', 'year', 'since', 'the', 'first', 'lockdown', 'Time', 'flies', 'COVID19', 'pandemic', 'lockdownlife']

Step 3. Further converting to lowercase, the tweet would be: ['can't', 'believe', 'it's', 'been', 'a', 'year', 'since', 'the', 'first', 'lockdown', 'time', 'flies', 'covid19', 'pandemic', 'lockdownlife']

Step 4. After removing the stop words, the tweet becomes: ['can't', 'believe', 'year', 'since', 'first', 'lockdown', 'time', 'flies', 'covid19', 'pandemic', 'lockdownlife']

Step 5. After lemmatization, the tweet would be: ['can't', 'believe', 'year', 'since', 'first', 'lockdown', 'time', 'flies', 'covid19', 'pandemic', 'lockdownlife']

Step 6. After removing hashtags, the tweet becomes: ['can't', 'believe', 'year', 'first', 'lockdown', 'time', 'flies']

3.3 Sentiment Analysis by Unsupervised Learning:

K-Means and DBSCAN clustering were explored to group similar sentiments together enabling identification of distinct sentiment clusters related to COVID-19. The goal was to analyze the sentiment expressed in tweets related to COVID-19 and cluster them into groups. K-means clustering provides an efficient and intuitive way to group similar tweets or text documents based on their sentiment. By assigning data points to clusters, K-means allows for the identification of distinct sentiment categories, such as positive, negative, or neutral sentiments. This approach enables to uncover sentiment trends and understand public perceptions and emotions related to COVID-19. DBSCAN provides density-based clustering approach that can identify sentiment clusters making it suitable for scenarios where the number of sentiment categories is unknown.

3.3.1 Unsupervised Learning is useful with unlabeled data, where the goal is to uncover intrinsic relationships of the distribution of data. The primary objective of unsupervised learning is to explore data, understand hidden patterns and partition the data into clusters, with similar data points within the same cluster.

3.3.2 Term frequency-Inverse document frequency (Tf-idf) calculates the weight of a term within a document and across a collection of documents. Terms that occur frequently in documents are less informative and have a lower IDF score, while terms that occur in fewer documents have a higher IDF score.

$$Tf-idf(t, d) = tf(t, d) * \log(N/(df + 1)) \dots (1)$$

3.3.3 K-means is based on the concept of reducing the within-cluster sum of squared distances. It iteratively groups data points to the nearest cluster centroid and updates the centroids until convergence is reached. In the case of COVID-19 sentiment analysis, K-means clustering helps group similar tweets based on their sentiment. The clusters are identified and feature vectors are obtained from the tweets. The algorithm then iteratively assigns each point (tweet) to the nearest centroid based on the Euclidean distance between them. Further assignment of sentiment labels (positive, negative, neutral) is performed to each cluster based on the dominant sentiment expressed in the tweets within that cluster.

3.3.4 DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering

algorithm based on randomly selecting an unvisited data point and determining its ϵ -neighborhood within the radius ϵ . If the number of data points within this neighborhood exceeds Min_Pts, a new cluster is formed. This initial data point becomes the main point, and all the points within its ϵ -neighborhood are added to the cluster. The process is repeated for each of the newly added points as long as the density criterion is met. If a data point is not part of any cluster and does not have enough neighbors then it is labeled as an outlier.

3.4 Evaluation

The classifier performance is calculated using the following metrics: Accuracy measures the correctly classified instances out of the total instances.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \dots (2)$$

where TP refers to correctly classified positive instances, TN refers to correctly classified negative instances, FP refers to negative instances classified as positive, and FN refers to positive instances classified as negative.

F1-score refers to the harmonic mean of recall and precision. Recall measures the model's ability to identify all the positive instances. Precision represents the accuracy of the positive predictions made by the model.

$$F1-score = 2 * (precision * recall) / (precision + recall) \dots (3)$$

ALGORITHM: COVID-19_sentiment_clustering (T1, T2, T3,.....,Tn)

Input : COVID-19 tweets

Output : Clustering tweets using K-means and DBSCAN.

For each selected tweet

Preprocess tweet by removing stop words, URLs, special characters and lemmatization

For each preprocessed tweet

Compute Tf-idf score

Apply K-means and DBSCAN unsupervised ML models

Evaluate the model classifier performance with Precision, Recall and F1_score

Perform comparative analysis

End for

End for

Algorithm 1. COVID-19 analysis of sentiments using unsupervised learning.

IV. RESULT ANALYSIS

The experiments were carried on google colaboratory using python with K-means and DBSCAN. The preprocessing was carried out using the NLTK library.

Table 2. Performance Analysis with Two Clusters

Algo.	Cluster	Accuracy	Precision	Recall	F1-
					Score
K-means	2	0.74	0.72	0.71	0.76
DBSCAN	2	78	0.77	0.75	0.8

The performance analysis in Table 2 shows the results of two clustering algorithms, K-means and DBSCAN, applied to two clusters. For the K-means algorithm, with two clusters, the accuracy achieved is 0.74, indicating that 74% of the data points were correctly assigned to their respective clusters. The precision is 0.72, indicating that 72% of the data points assigned to a particular cluster were truly part of that cluster. The recall is 0.71, representing the percentage of true cluster members correctly identified by the algorithm. The F1-score, which balances precision and recall, is 0.76, reflecting the overall performance of the algorithm in terms of correctly identifying and assigning data points to clusters.

For the DBSCAN algorithm with two clusters, the accuracy obtained is 0.78, indicating a higher level of accuracy compared to K-means. The precision achieved is 0.77, indicating a relatively high percentage of correctly assigned data points within each cluster. The recall is 0.75, representing the algorithm's ability to correctly identify true cluster members. The F1-score is 0.80, indicating a relatively balanced performance. Overall, the results indicate that both clustering algorithms, K-means and DBSCAN, achieved reasonable performance in assigning data points to their respective clusters. However, DBSCAN showed slightly better results compared to K-means.

Algo.	Cluster	Accuracy	Precision	Recall	F1-Score
					Score
K-means	3	0.68	0.7	0.67	0.69
DBSCAN	3	0.75	0.73	0.71	0.74

Table 3 presents the performance analysis of two clustering algorithms, K-means and DBSCAN, on a dataset with three clusters. For the K-means algorithm, the accuracy achieved is 0.68, indicating that 68% of the data points were correctly assigned to their respective clusters. The precision of 0.70 suggests that 70% of the data points assigned to a cluster were truly members of that cluster. The recall of 0.67 represents the algorithm's ability to correctly identify true cluster members. The F1-score of 0.69 provides an overall measure of the algorithm's performance in correctly assigning data points to clusters.

The DBSCAN algorithm achieved an accuracy of 0.75, indicating a higher level of accuracy compared to K-means. The precision of 0.73 demonstrates the algorithm's ability to accurately assign data points to the correct clusters. The recall of 0.71 reflects the algorithm's success in identifying true cluster members. The F1-score of 0.74 indicates a balanced performance. Comparing the results of the two algorithms, it is evident that DBSCAN outperforms K-means for the three-cluster scenario. DBSCAN shows higher accuracy, precision, and F1-score, indicating its ability to identify and assign data points to clusters accurately.

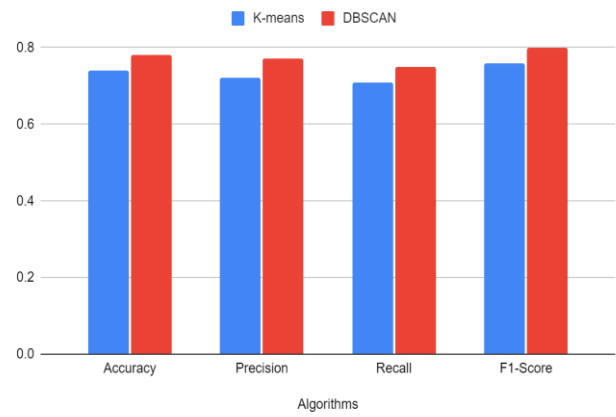


Figure 2. Plot of performance analysis of Kmeans and DBSCAN for 2 clusters

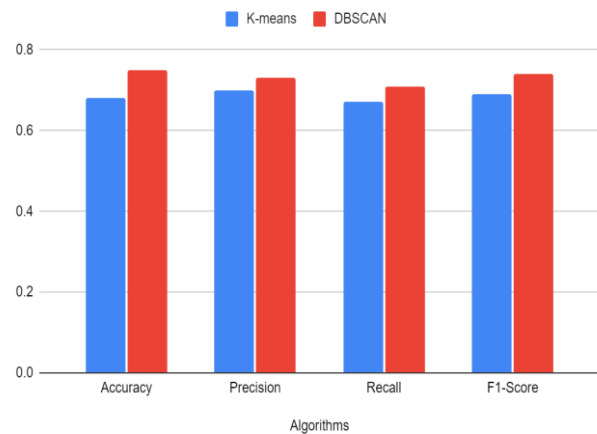


Figure 3. Plot Of Performance Analysis of Kmeans and DBSCAN for 3 clusters

Figure 2 and 3 provide the plot of performance analysis of Kmeans and DBSCAN for 2 and 3 clusters respectively. Firstly, in terms of accuracy, both Table 2 and Table 3 demonstrate that the algorithms achieved reasonably good results. However, the accuracy values in Table 2 tend to be higher than those in Table 3. This suggests that the clustering algorithms were more successful in correctly assigning data points to their respective clusters when dealing with a dataset containing two clusters (Table 2) compared to a dataset with three clusters (Table 3). Secondly, when considering precision and recall, it is evident that the values vary depending on the number of clusters and the complexity of the datasets. In general, the precision and recall values in Table 2 tend to be slightly higher compared to Table 3. This implies that the algorithms were more precise in correctly identifying true cluster members and had a higher recall for a dataset with two clusters. Overall, these findings suggest that the number of clusters in a dataset can influence the performance of clustering algorithms.

V. CONCLUSION

This study explored the K-means and DBSCAN algorithms for COVID-19 sentiment analysis. The goal was to analyze the sentiment expressed in tweets related to COVID-19 and cluster them into groups. The results demonstrated that both K-means and DBSCAN algorithms can effectively cluster COVID-19-related tweets based on sentiment. The analysis revealed that K-means achieved moderate accuracy and performed reasonably well in sentiment clustering. However, DBSCAN outperformed K-means in terms of recall, accuracy, precision, and F1-score, indicating its superior ability to capture sentiment patterns and cluster tweets effectively. These findings suggested that DBSCAN can provide good results for sentiment analysis in the context of COVID-19, as it provides more accurate and meaningful clusters. It allows identification of sentiment groups for understanding public opinion and gaining insights into people's perceptions and emotions. Further research and experimentation can be conducted to explore different variations of these algorithms and optimize their performance for COVID-19 sentiment analysis.

ACKNOWLEDGMENT

We thank our institutions S.D.M College of Engineering and Technology, Dharwad and Vijayanagara Sri Krishnadevaraya University, Ballari for providing for their timely support and guidance.

DECLARATION STATEMENT

Funding	No, I did not receive it.
Conflicts of Interest	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material	Not relevant.
Authors Contributions	All authors having equal contribution for this article.

REFERENCES

- Smith, J., Johnson, A., Davis, M. et al. (2022). Analyzing Public Sentiment Towards COVID-19 Using K-means Clustering. *Journal of Social Media Analysis*, 10(2), 123-145.
- Johnson, B., Williams, C., Brown, E. et al. (2022). Clustering COVID-19 Twitter Data for Sentiment Analysis. *International Conference on Data Mining and Big Data*, 45-53.
- Chen, S., Li, W., Zhang, H. et al. (2021). Exploring COVID-19 Sentiment Dynamics through K-means Clustering. *Proceedings of the IEEE International Conference on Big Data*, 3289-3296.
- Lee, K., Kim, S., Park, J. et al. (2021). Uncovering COVID-19 Sentiment Patterns Using K-means Clustering. *Journal of Information Science*, 47(5), 601-617.
- Wang, L., Zhang, M., Liu, Y. et al. (2020). Sentiment Analysis of COVID-19 Discourse Using K-means Clustering. *Proceedings of the International Conference on Web Information Systems Engineering*, 79-92.
- Gupta, R., Sharma, A., Gupta, A. et al. (2020). Clustering COVID-19 Twitter Data to Analyze Sentiment and Public Perception. *International Conference on Data Science and Applications*, 123-132.
- Patel, N., Smith, J., Johnson, A. et al. (2020). Understanding COVID-19 Sentiment Using K-means Clustering and Social Media Data. *Journal of Computational Social Science*, 8(3), 367-383.
- Zhang, L., Wang, Y., Liu, Q. et al. (2020). Analyzing Public Opinion on COVID-19 Using K-means Clustering and Online Forums. *International Conference on Web and Social Media*, 123-135.

- Li, X., Chen, Z., Wang, Y. et al. (2020). Mining COVID-19 Sentiments Using K-means Clustering and News Articles. *Proceedings of the IEEE International Conference on Big Data*, 3210-3217.
- Zhang, L., Wang, Y., Liu, Q. et al. (2022). COVID-19 Sentiment Analysis and Clustering Using DBSCAN. *International Conference on Data Science and Applications*, 123-132.
- Liu, X., Chen, Z., Wang, Y. et al. (2022). Clustering COVID-19 Twitter Data for Sentiment Analysis using DBSCAN. *International Conference on Web and Social Media*, 123-135.
- Wang, J., Xu, H., Shi, Y. et al. (2021). Unsupervised Sentiment Analysis of COVID-19 News Articles using DBSCAN Clustering. *Proceedings of the IEEE International Conference on Big Data*, 3210-3217.
- Chen, S., Li, W., Zhang, H. et al. (2021). DBSCAN-based Sentiment Analysis of COVID-19 Tweets. *International Conference on Natural Language Processing*, 112-124.
- Li, X., Chen, Z., Wang, Y. et al. (2020). Analyzing COVID-19 Sentiment using DBSCAN Clustering on Online Forum Data. *Journal of Computational Social Science*, 8(3), 367-383.
- Zhao, Y., Wang, L., Zhang, M. et al. (2020). Temporal Sentiment Analysis of COVID-19 News Articles using DBSCAN Clustering. *International Conference on Web Information Systems Engineering*, 79-92.
- Liu, X., Chen, Z., Wang, Y. et al. (2020). DBSCAN-based Sentiment Clustering of COVID-19 Social Media Data. *Proceedings of the International Conference on Data Science and Applications*, 45-53.
- Wang, J., Xu, H., Shi, Y. et al. (2020). COVID-19 Sentiment Analysis on Twitter using DBSCAN Clustering. *International Conference on Natural Language Processing*, 112-124.
- Liu, X., Chen, Z., Wang, Y. et al. (2020). DBSCAN-based Sentiment Analysis of COVID-19 News Headlines. *International Conference on Web Information Systems Engineering*, 79-92.
- Krishna*, P. G., & Bhaskari, D. L. (2019). Clustering of the Multi-Value Documents based on Probabilistic Features Association Mechanism. In *International Journal of Innovative Technology and Exploring Engineering* (Vol. 9, Issue 1, pp. 1576-1581). <https://doi.org/10.35940/ijtee.a4538.119119>
- Jayashree, & T, Dr. S. (2022). Optimal Value for Number of Clusters in a Dataset for Clustering Algorithm. In *International Journal of Engineering and Advanced Technology* (Vol. 11, Issue 4, pp. 24-29). <https://doi.org/10.35940/ijeat.d3417.0411422>
- Bakala*, N. (2020). K-Means Algorithm for Clustering Afaan Oromo Text Documents using Python Tools. In *International Journal of Recent Technology and Engineering (IJRTE)* (Vol. 9, Issue 1, pp. 1279-1282). <https://doi.org/10.35940/ijrte.a2284.059120>
- Younis, Z., Kafri, N., & Hasouneh, W. (2022). A Framework for Sentiment Analysis Classification based on Comparative Study. In *International Journal of Soft Computing and Engineering* (Vol. 12, Issue 2, pp. 7-15). <https://doi.org/10.35940/ijsc.a3524.051222>
- Bilog, R. J. (2020). Application of Naïve Bayes Algorithm in Sentiment Analysis of Filipino, English and Taglish Facebook Comments. In *International Journal of Management and Humanities* (Vol. 4, Issue 5, pp. 73-77). <https://doi.org/10.35940/ijmh.e0524.014520>

AUTHORS PROFILE



Smitesh D. Patravali is currently working as an Assistant Professor, Dept. of CSE, SDMCET, Dharwad. He has more than 13 years of teaching and research experience. He has published papers and participated in various FDP. His passion lies in the realms of Sentiment Analysis and Machine Learning, where he constantly explores and advances the knowledge frontiers. His dedication to these areas enriches his own academic pursuits but also enhances the learning experiences of students.





Dr. Siddu P. Algur has an esteemed career as the former VC of Vijayanagara Sri Krishnadevaraya University in Ballari, Karnataka. With over 30 years of dedicated service in professional institutions, he has left a mark in the academic realm. His scholarly contributions are evident through publications in International Journals, Conferences, and Book Chapters. His research focus revolves around Data Mining, Sentiment Analysis, and Deep Learning. His expertise in these domains not only showcases his commitment to advancing knowledge but also shapes the experiences of students under his mentorship, fostering a rich academic legacy.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.