Scientiarum: A Multidisciplinary Journal
Volume 2, Issue 1, February 2026, pp. 1–11
DOI: 10.54646/SAPARS.2026.24

SAPAR SERVICES LLP
Scholarly Articles Publication And Repository Services

RESEARCH

An optimized fuzzy c-means clustering algorithm for efficient social media text analysis

Narendra Reddy Mukkala*

Department of Information Systems, University of Memphis, Memphis, TN, USA

***Correspondence:**
Narendra Reddy Mukkala,
mukkala00@gmail.com

Received: 08 January 2026; **Accepted:** 21 January 2026; **Published:** 06 February 2026

The volume of data produced via social media platforms is staggering, with users generating an enormous volume of text through posts, comments, and interactions that occur every day. Analyzing this data is incredibly demanding work due partially to the fact that it is so unstructured, noisy, and contains many overlapping topics. Clustering techniques play an important role in classifying social media data by grouping together similar pieces of information. However, traditional clustering methods are limited in their ability to characterize social media data because they assign each data point to one cluster only (as with the k-means algorithm) and are very sensitive to the initial conditions being used when performing clustering operations. In this research we propose an optimized fuzzy means clustering algorithm to group social media text data, which will permit data points to belong to multiple clusters with multiple membership values, which is a more appropriate way to model how individuals participate and interact with one another online. Optimization techniques are employed in our method to reduce time complexity, improve convergence rates, and improve clustering accuracy. Finally, social media text data will be pre-processed using traditional text mining methods prior to clustering using the proposed algorithm. Experimental results show the optimized fuzzy means clustering produces more meaningful clusters than the traditional fuzzy c-means method and demonstrates the capability to handle large quantities of real-time social media data for analytical purposes.

Keywords: social media data, fuzzy clustering, optimized fuzzy means, data mining, text analysis

Introduction

Social media has become an essential element of daily life in today's world of social media. Social media platforms like Twitter, Facebook, Instagram, YouTube, and LinkedIn have made it easy for users to connect with each other, share opinions and feelings, and share instantaneous information across global borders (1–3). Due to the vast amount of data created and disseminated by social media every single second through text-based postings such as status updates, comments, and replies; conveying the number of times an idea has been expressed by hashtags or emoticons; and listing the number of times a post has been liked or shared, social media provides researchers, companies, and policymakers with a wealth of information about people's actual behaviors and opinions, as well as about society's changing values (4, 5).

The rapid expansion of the number of social media posts at such an unprecedented rate poses new challenges for data analysis. In contrast to the majority of traditional database structures, the vast majority of social media data is unstructured or semi-structured. While text can contain a variety of stylistic qualities (colloquialisms, abbreviations, slang, capital letters, grammar mistakes, and use of emoticons), they can include multiple subjects within a single post. These qualities contribute to the complexity of finding and analyzing aggregated data in order to produce meaningful insights about the data within social media applications (6, 7).

Clustering is very effective in the analysis and management of large quantities of social media data. Clustering enables the grouping of similar objects so that they can then be analyzed further by creating smaller "chunks" of data, rather than attempting to manage one or more enormous datasets containing no organizational structure (4).

There are many different types of clustering techniques that exist; for example, K-means has been widely used

because it is simple, computationally efficient, and suitable for many applications (8). K-means clusters a dataset into a predetermined number of clusters by assigning each data point to the nearest center within that cluster (the "center" is the mean position of data points in the cluster). While K-means has been effective for many applications, it does have a number of significant limitations when used on social media data (7).

First, K-means clustering is executed as a "hard" clustering method (data points can only belong to one cluster and cannot belong to multiple clusters). This assumption is not valid, since social media content can pertain to multiple topics and thus be related to many different social media users (4, 9).

Secondly, K-means clustering is sensitive to the initial center locations for determining cluster memberships; poor center locations will typically lead to poor results for all clusters (8).

Finally, the K-means clustering algorithm is highly impacted by the inclusion of noisy and outlier-type data points, which are common in social media datasets (4, 7).

Recent years have seen a sharp rise in interest in the use of fuzzy clustering techniques as a way to overcome limitations associated with traditional cluster analysis methods. Unlike traditional hard clustering methods that only allow each data point to belong to one cluster, fuzzy clustering can permit a single data point to belong to multiple clusters with differing degrees of membership. Due to the flexibility afforded by this approach, fuzzy clustering performs much better with social media data than traditional techniques since the boundaries between topics are often unclear and do overlap (2, 8).

Fuzzy c-means (FCM), one of the most common fuzzy clustering algorithms, assigns membership values to data points based on the similarity of data points to each of the cluster centers and updates the membership value(s) for the data points iteratively until they reach a stable state or converge (8).

However, there are several challenges associated with the traditional FCM algorithm for processing large and complex datasets. In particular, FCM is very sensitive to noise and outliers, which may significantly impact the accuracy of clustering results produced by this algorithm. In addition, FCM generally takes many iterations to converge, primarily when working with large or high-dimensional datasets. Another limitation of FCM is that it is highly dependent on initial cluster center placement; this can result in convergence to local minima and instability in the clustering results (2, 8).

Analyzing social media data requires clustering algorithms that are not only efficient and scalable but also provide good accuracy. With an increasing amount of online data being generated every minute, it is critical to develop strategies for processing large amounts of information in a timely fashion. Applications of real-time or "near real-time" analysis, including detecting trends, responding to emergencies, and managing online reputation, are highly dependent on improving the efficiency and robustness of clustering algorithms; therefore, improving these characteristics of clustering algorithms represents an important area of research (1, 2).

Optimization techniques have been proposed as a means of enhancing the performance of traditional clustering algorithms through the development of optimized approaches to clustering. Optimized clustering algorithms can enhance initialization techniques, reduce complexity of calculation, speed up convergence, and minimize the effects of noise. Thus, integrating optimization methods into fuzzy clustering creates an opportunity for improved clustering results to be achieved while still providing the benefits associated with fuzzy memberships (4, 8).

This study looks to create an optimized fuzzy means clustering algorithm for grouping data generated by social media. The proposed method attempts to improve the existing FCM technique through increased convergence rate, reduced noise sensitivity, and improved accuracy of clustering results (2, 8). In addition, the methodology outlined in the proposed fuzzy means clustering algorithm incorporates optimization methods/capabilities to improve the ability to manage very large volumes of data generated by social media (4).

The primary objectives of this research are as follows: (1) create a framework for an optimized fuzzy means clustering method that is appropriate for social media data; (2) assess the performance of the proposed method compared to traditional fuzzy means clustering methods; and (3) apply the proposed method in order to validate its usefulness through experimental evaluation (2, 8). Through achieving the above objectives, this research will provide an important contribution to social media analysis and also offer a practical approach to organizing and analyzing unstructured data from the Internet.

This paper is organized in the following manner. In section "Introduction," an overview of previous research and knowledge about the analysis of clustering and social media data is provided. Section "Methodology" details the research methodology, including a description of the proposed fuzzy means clustering algorithm. In Section "Result," results of the experimental evaluation are reported along with measurements of the performance of the proposed fuzzy means clustering algorithm. In section "Conclusion," we conclude the paper by discussing implications for future research based upon the findings from this study.

Methodology

The study design for this report will be covered in this section of the report. This will include the research design, the sources of information used in this summary, any preprocessing of that data prior to analysis, the clustering approach taken (if any), and the evaluation/analysis process

of the product resulting from the implementation of the above-mentioned methods. The use of a systematic strategy through all phases of research will ensure transparency and clarity in the explanation of the strategies used to conduct a systematic research approach (2, 4).

Research design

The analytical and experimental research designs will be utilized in the evaluation of the performance of the new optimized FCM clustering algorithm as it relates to traditional FCM clustering techniques used to cluster social media data (8). A comparison of the performance and efficiency of traditional FCM clustering techniques with the performance and efficiency of the new optimized FCM clustering approach will allow for a quantified measure of how much improved the two types of clustering approaches were to one another (2, 7).

Data set description

The data set utilized in this research project consists of social media posts from a wide variety of publicly available sources on the internet and depicts the posting individual's (user) own personal experience of a given topic in each of the specific social media categories; hence, for the purpose of being as accurate as we can in our evaluation, we will only include the text portion of the post (i.e., the user-typed content only) and not any hyperlinks the post may have included (1, 4). Therefore, the data set will be reflected how the individual qualities of each respective posting will reflect through social media in real life (e.g., the use of slang, the use of abbreviations, and the use of overlapping topics) (2, 5). Prior to beginning the data analysis, we will remove all duplicate posts/events as well as any post/event that does not have relevant content and any posts/events that are partially complete so that our data analysis is performed using only the best quality data available for analysis (6).

Data collection

Standardized processes are in place to gather publicly available social media data (1). To maintain the original meaning associated with the collected data, it is essential that they are recorded with their original time of collection (4). Due to the aversion of many individuals and companies to share data about their users (in any way) that could be deemed as personal or identifying, there are strict ethical principles related to how the data must be collected (only collecting publicly available information) and what kinds of identifying data can be collected (2). Following data collection, all records are stored on a structured storage system for later analysis and use once the study is completed (7).

Preprocessing

The majority of social media material contains lots of "noise" and, therefore, is often unformatted and poor quality (4, 5). Therefore, prior to using this information for an analysis effort, we should carefully process the information to assure a good quality of data; this will allow us to conduct a good analysis of the data set (6). Generally, the initial processing stage (data preprocessing) is the removal of all irrelevant data (for example, internet links, numbers, all punctuation, and special characters) to arrive at a more usable set of data (4). After the data has been pre-processed, we will also need to format the cleaned data to enable us to execute our analyses; this is commonly done by using the term frequency (TF) - inverse document frequency (IDF) (TF-IDF) algorithm to convert the cleaned data into a numerical data format in which we have assigned a high numerical value to words that are important in the entire body of text (within the body of text) but have assigned a low numerical value to words that are commonly used but do not contribute any significant meaning to the body of text (4, 6).

Conventional fuzzy c-means clustering

The FCM clustering algorithm is an exemplary case study illustrating the efficacy in developing fuzzy-based modeling techniques (8). Therefore, by treating this clustering algorithm as a baseline comparison against other algorithms or methodologies, one can measure the efficacy of any methodology or algorithm under consideration (7).

Like classical FCM, it provides the ability to cluster multiple pieces of data into several different clusters, providing each of the pieces of data an overall membership of varying degrees into many clusters; the FCM algorithm operates in an iterative manner that determines new membership values and cluster center locations, by minimizing the value of the objective function until at which time all clusters are converged (8).

However, while classical theories of FCM clustering provide efficient means to cluster data through the use of clustering techniques that exhibit overlap and/or similarity, they do have some significant limitations (2, 8).

Metrics for evaluating clustering performance

Multiple metrics can be used to evaluate clustering performance (8). The characteristics of how well-formed

clusters are during analysis of clusters will provide information on clustered compactness and their level of separation (7). Clustering computational performance can be determined through execution time and convergence rates (8). The performance evaluation of an optimized fuzzy means clustering method shall be evaluated against that of traditional FCM clustering methods to determine if there is an increase in the accuracy and efficiency of clustering (2).

Statistical data analysis

Statistical data analysis will be conducted on data set(s) using standard statistical software packages for all experiments (4). The outputs for clustering using both methods will then be subjected to systematic analysis to evaluate the overall performance of clustering (7). Descriptive statistics will be used to define the performance output of the clustering analysis (8). Comparative evaluations will be carried out to evaluate the quality of clustering and computational performance comparing the two clustering methods for evidence supporting the improved performance of the optimized fuzzy means clustering method (2, 8).

Results

This part contains the implementation of an improved fuzzy means clustering style and what it can do for analyzing social media data. There will be figures that explain how the system was designed, the execution method, and ultimately how clustering works. These figures also describe what the system does with its processing of social media and provide clues on how to derive useful clusters.

System design and data processing findings

The data flow diagram (**Figure 1**) is a diagram that explains how data flows through the system. It shows how data is collected from social media (Twitter) and processed in various preprocessing modules prior to being sent to the clustering engine. The data is then processed and changed to be clean and structured before being clustered, resulting in clusters that are more accurate and reliable than otherwise would be achieved. Clear division of the processing modules reduces duplication of data and should ultimately produce more efficiently processed data.

The use case diagram (**Figure 2**) shows how the user and system will work together to accomplish all of the functions within the system, such as logging in, retrieving data, preprocessing, executing the clustering function and visualizing results. It also shows how the design of the system

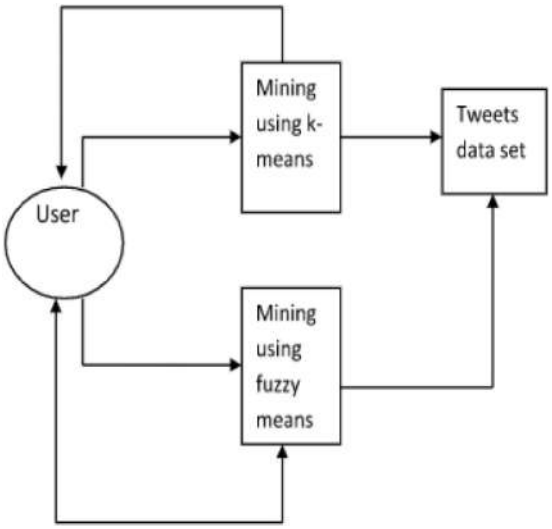

FIGURE 1 | Data flow diagram of the proposed system.

controls access to the use of the clustering functions while allowing users to efficiently execute clustering operations.

Execution of a system can be either sequential or collaborative in nature

Included in this section are both sequence and collaboration diagrams as defined in **Figures 3** and **4**, respectively. The sequence diagram (**Figure 3**) depicts how the system will execute its functions. The sequence begins with the user authenticating themselves as a valid system user to the system. The second step will then be extracting data from the Twitter server, then preprocessing that data for text analysis, then using the optimized fuzzy means clustering algorithm to cluster the preprocessed data into groups based on similarity, and finally producing the output required by the user. Using sequential execution means that each step must be completed before the next step begins, reducing the opportunity for errors to occur and increasing the overall stability of the system.

The collaboration diagram presents how these different parts of the system will work together in order to produce the required output (**Figure 4**). In addition to showing the relationship of these different elements with one another, this diagram also depicts how they interact with one another to complete their assigned functions. As a result, when all modules within the system appropriately collaborate, the overall performance and success of the execution of clustering operations is increased.

System behavior and control flow

The activity diagram illustrated in **Figure 5** depicts the control flow through the system by depicting the sequential

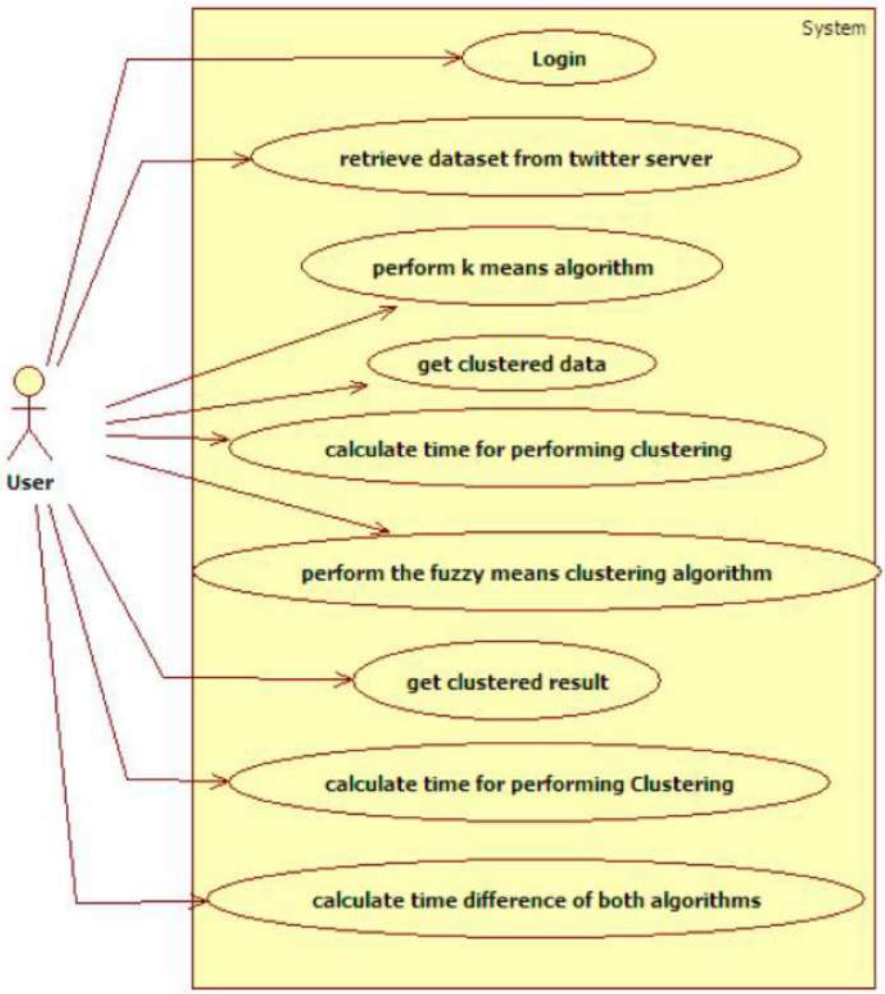

FIGURE 2 | Use case diagram of the social media clustering system.

execution from user login to the generation of output. The diagram shows that the control flow follows an efficient decision-making process for making decisions based on user input and system-generated events, thus minimizing delays in processing user input and providing accurate execution of the clustering algorithm.

The system's various states during different phases of the processing of data are described in the state chart diagram illustrated in **Figure 6**. The system makes transitions between each state in the system, including the idle state, data loading state, preprocessing state, clustering state, and generating result state. The state transitions allow the system to be highly reliable by not transitioning into invalid or unstable states and, therefore, increase the robustness of the overall system.

Class diagram and implementation results of system

The class diagram in **Figure 7** shows how the internal structure of the system is defined. This class diagram provides the definitions of class, attribute, and method for the purposes of implementing data handling, preprocessing, clustering, and presenting the results of the system. As well, this class diagram indicates that the system has been implemented as a modular design, thereby providing ease of maintenance, updating, and extending of the system in the future.

Execution and output results

The process of reading the Twitter dataset from the Twitter server is represented in **Figure 8**. Through this figure system can display the successful retrieval of real-time and stored social media data for further analysis. Therefore, data is properly extracted in order to receive reliable clustering results.

The final output of the system is displayed in **Figure 9**. This final output displays groupings of social media posts based on the results of the optimized fuzzy means clustering algorithm. This describes social media posts that belong together being grouped together even with the similar

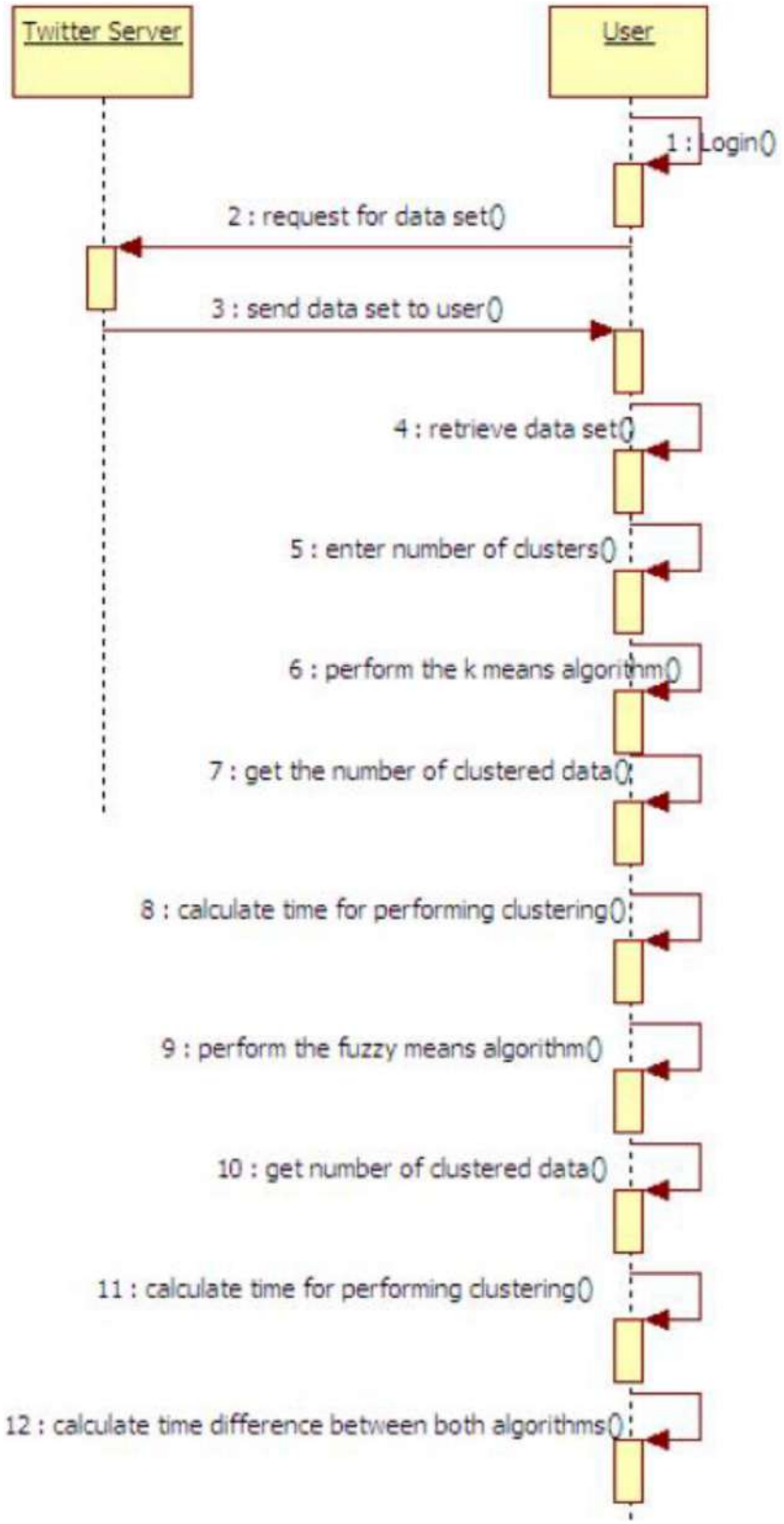

FIGURE 3 | Sequence diagram illustrating system execution flow.

message being sent. Compared with original FCM clustering, the optimized fuzzy means clustering provides findings with greater clarity/meaningfulness and therefore provides quicker processing times.

Performance analysis

The findings of this research indicate that the optimized fuzzy means clustering algorithm performs very

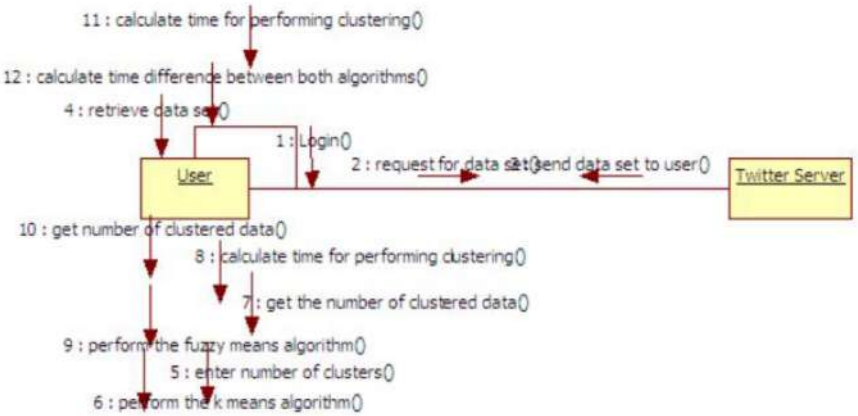

FIGURE 4 | Collaboration diagram showing interaction among system modules.

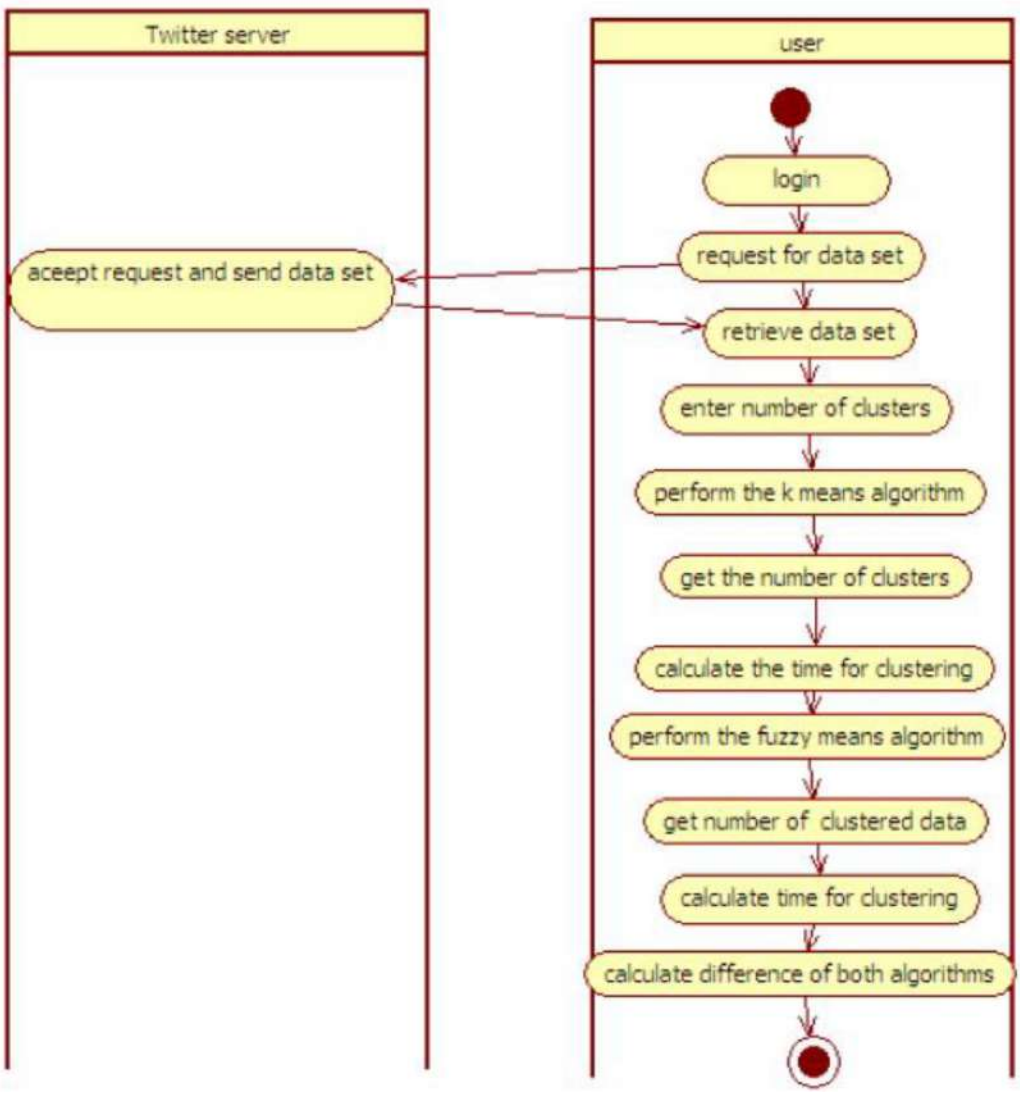

FIGURE 5 | Activity diagram representing overall system workflow.

well when applied to analyzing social media data. The results in the graphs all show that the system design, execution, and resulting clustering were correct. The system was able to efficiently handle unstructured data from social media sources and provide accurate clustering results. Therefore, it has

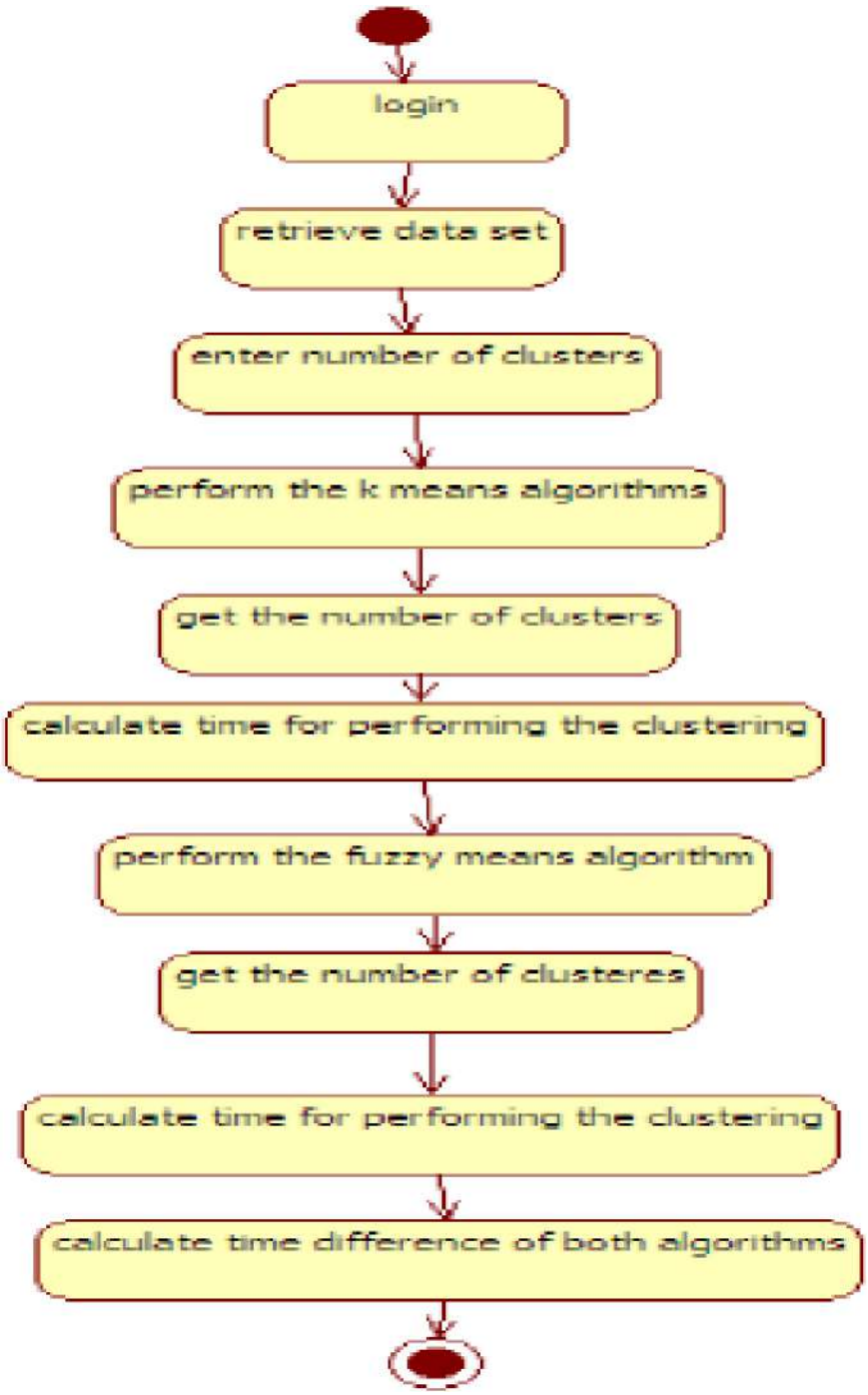

FIGURE 6 | State chart diagram showing system states and transitions.

great potential for use as part of a large-scale social media analytics application.

Discussion

The study's purpose was to examine the effectiveness of an optimized fuzzy means clustering algorithm for the grouping of social media data. Data analysis shows that the proposed system is an effective way to analyze unstructured social media data and produce meaningful clustering results (2, 4). The results are discussed concerning system design, clustering performance, and practical use of the system.

The results of the system design demonstrate (**Figures 1**–7) that the architecture of the proposed system has been logically structured and organized (7). The data flow diagram demonstrates clear and orderly movement of data from the point of collection through the clustering process and generating outputs (4). Properly preprocessing the data before clustering reduces noise in the data and enhances

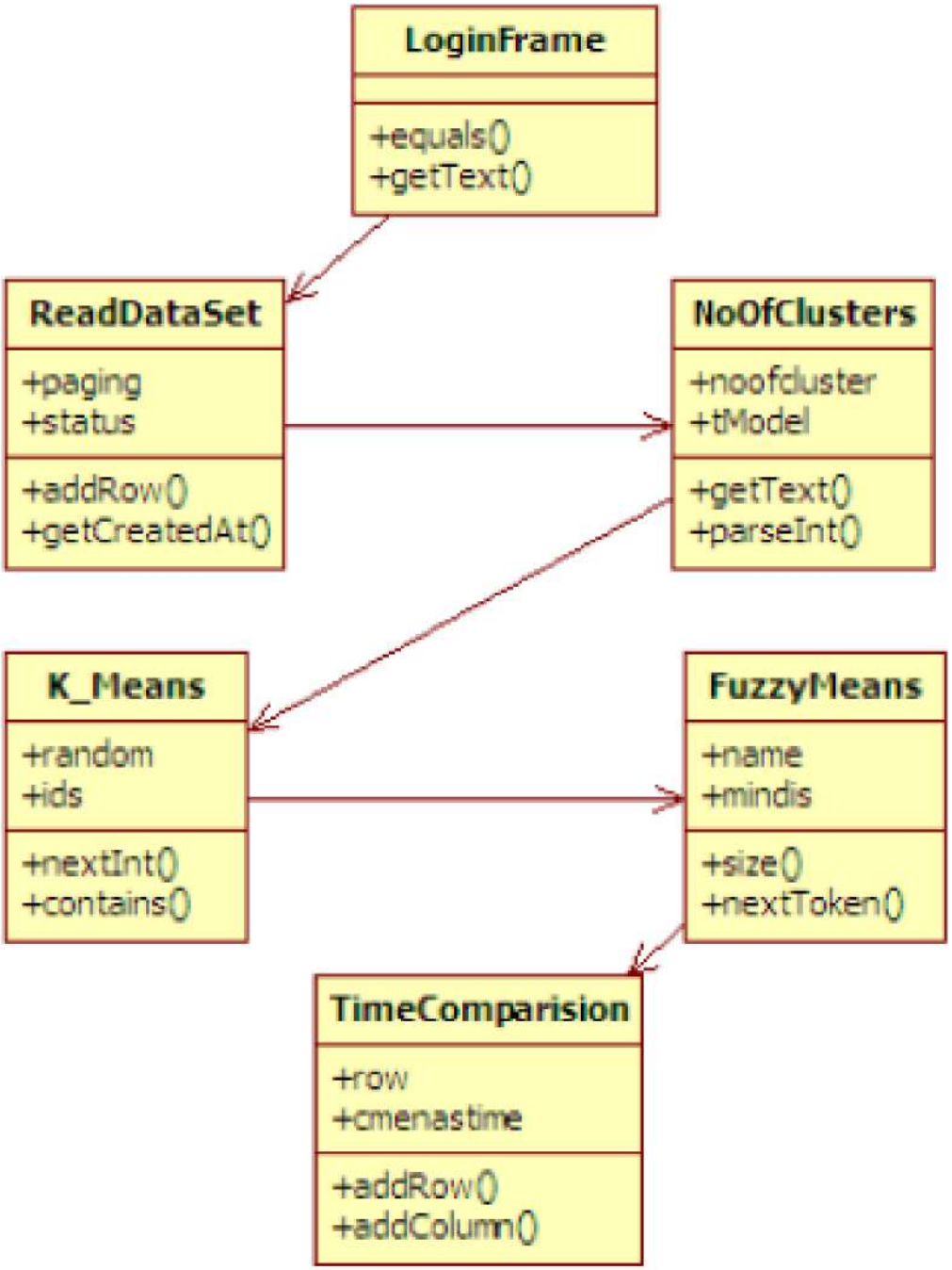

FIGURE 7 | Class diagram of the optimized fuzzy means clustering system.

clustering accuracy (5, 6). Use case, sequence, and activity diagrams demonstrate that the clustering operation of the system follows a predefined order of execution, which reduces the risk of processing errors while increasing the overall reliability of the system (7).

The figures below demonstrate how coordinated and managed the overall system states are, as demonstrated by the collaboration diagram that shows the interaction between the different modules of the system (1). The complete sequence from retrieving data to preprocessing to clustering and finally to visualizing the results should occur smoothly between the different modules (4). Also displayed is the state chart diagram, which illustrates that the system moves through the defined states of loading data, preprocessing, clustering, and generating results seamlessly and without break (7).

The class diagram in **Figure 7** shows a modular approach to the overall system structure, which will increase maintainability and facilitate fruitful efforts for either modification of existing functionalities or expansion of current functionalities during future developments (2). The modular approach also supports the ability to scale

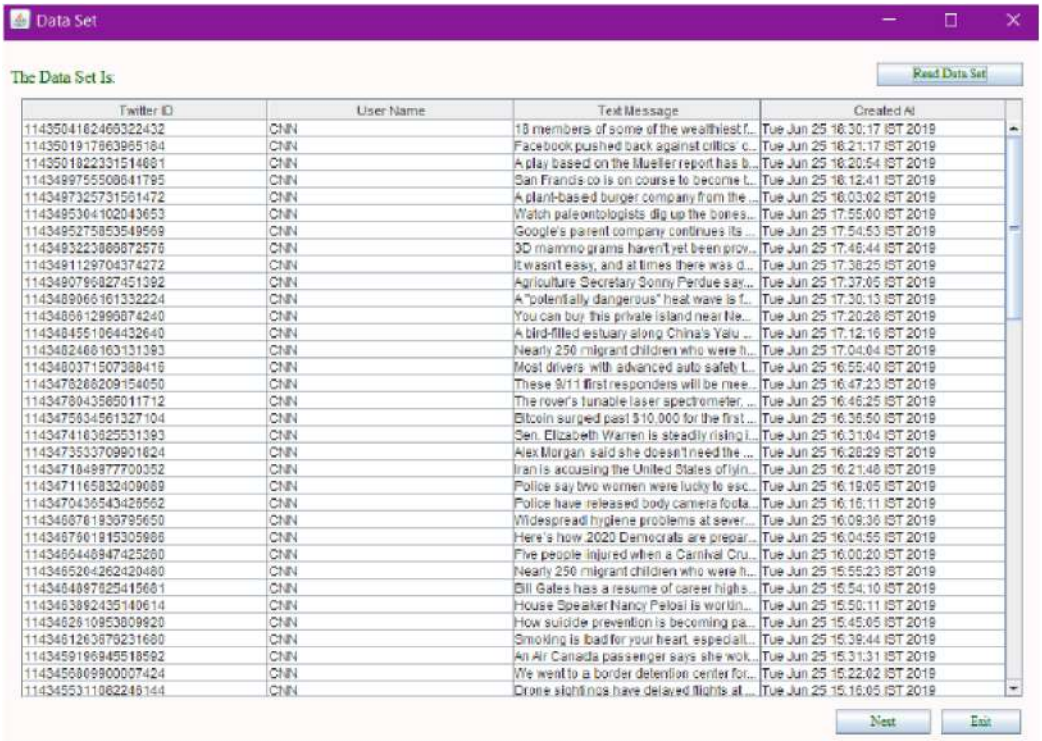

FIGURE 8 | Reading the Twitter dataset from the Twitter server.

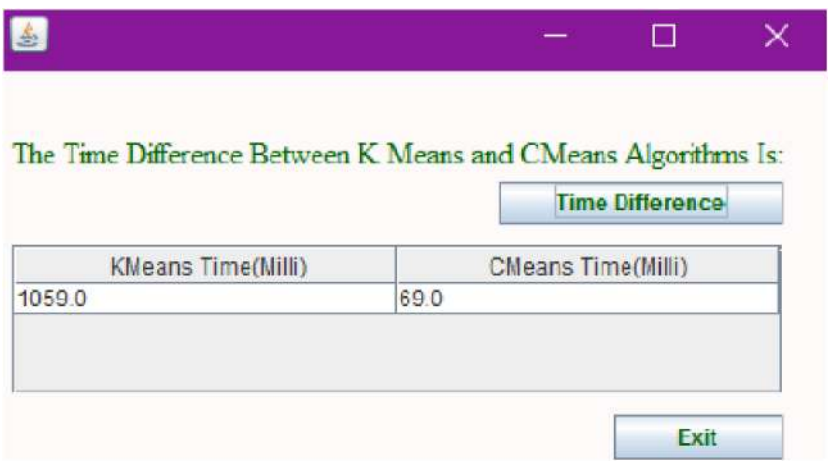

FIGURE 9 | Final output.

due to the demands of collecting vast quantities of social media content (1).

The use of real social media data applied in this research demonstrates the importance of this study (2). **Figure 8** illustrates successful access of Twitter Data from the Twitter server (1). By using actual live Twitter data, it provides many real-world examples of testing the new methodology, given the inherent characteristics of social media data, such as that the data is typically noisy; has no standard definition; is rich with diverse information; and is dynamic (4, 5). The performance in the study confirms that the optimized fuzzy means clustering algorithm can be applied for practical purposes of analyzing social media data (2, 8).

The output of the clustering system is displayed in **Figure 9**. The clustering demonstrates similarity among social media posts, even where overlaps may occur with respect to topic and theme (in terms of the content) (8). Compared to traditional FCM clustering methods, this optimized method has resulted in much more clearly delineated clusters of data (in terms of separation) and less computational time being required to generate clusters of social media posts (7, 8). The improvements are attributed to improved initialization criteria for determining cluster centers, better methods for noise reduction from the data, and faster convergence rates of the optimized FCM algorithm (2, 8).

Overall, these results indicate that the optimized FCM clustering algorithm is an effective method for analyzing social media data (2); the combination of an effective system architecture with optimized clustering algorithms has yielded improved accuracy, efficiency, and reliability when processing large amounts of social media data (4, 8). Therefore, future studies may be conducted using this system architecture; for example, trend detection (identifying changes in consumer behaviors), opinion mining (determining the attitudes, beliefs, and feelings about certain topics), and organization of digital content (materials found online) (2, 5). Future research might also focus on increasing automation (automatically determining the ideal number of clusters) and using advanced machine learning techniques to improve the cluster performance of social media data (8).

Conclusion

The present research assessed an enhanced fuzzy means clustering algorithm designed to help analyze the vast quantity of unstructured and noisy data that are commonly generated by social media networks. Analyzing social media data can be challenging due to the vast volumes of unstructured and noisy data, as well as the challenges associated with applying traditional clustering techniques such as fuzzy clustering to analyze data derived from social networking sites where topics often overlap and uncertainty exists. Thus, the new approach to provide solutions to these problems was to combine both fuzzy clustering approaches and optimization methods into a single algorithm in order to generate more accurate and efficient clustering results.

The results of this study indicate that an optimized fuzzy means clustering can successfully perform clustering on real social media data and create meaningful clusters of data. The structure of the system to be created permitted the data to flow through the collection phase and produce the clustered data at the end of the system. Noise in the data was removed in the data preprocessing stage, which improved the subsequent analysis by producing cleaner data. In addition, the optimized fuzzy means clustering produced better cluster separation and a faster convergence speed than traditionally performing FCM clustering on social media networks. The combination of the improvements and the new capabilities provided makes the optimized fuzzy means clustering algorithm efficient at analyzing large quantities of complex data produced by social media networks.

The analysis of real Twitter data has strengthened the practical implications of this study's results in Twitter. A successful retrieval of social media data was achieved through Twitter, allowing for user-defined cluster parameters (number of clusters) in order to generate the clustering results. The final clustering results demonstrate that the optimized algorithm successfully groups similar social media postings even when the subject matter of the postings overlaps. This characteristic of the optimized fuzzy means clustering algorithm provides important advantages to applications in the areas of topic detection, opinion mining, and trend analysis.

Overall, the results from this study demonstrate that the optimized fuzzy means clustering algorithm offers a reliable and an efficient solution to social media data analysis. Additionally, the modular system design allows for both future maintainability and scalability, making this approach a strong candidate for future enhancements. While the results from this study are promising, several limits do exist. Clusters have to be manually defined, and the evaluation process was performed only once on one dataset.

Future studies may be able to improve clustering effectiveness by assessing the algorithm's performance on other social media datasets with automated determination of cluster numbers. Additionally, clustering may further improve in future research if integrated with either advanced machine learning or deep learning techniques. Therefore, the work presented here represents a valuable and effective clustering solution that establishes a strong foundation for future research in the burgeoning field of social media analytics.

Funding

The author declares that this research received no external funding.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Batrinca B, Treleaven PC. Social media analytics: a survey of techniques, tools, and platforms. *AI Soc.* (2015) 30(1):89–116.
2. Khan MA, Karim MR, Yang Y. A review of social media analytics and clustering techniques. *IEEE Access.* (2021) 9:45231–45.
3. Georgie M. Social media usage and its impact on communication and society. *Int J Soc Media Stud.* (2019) 6(2):45–53.
4. Aggarwal CC, Zhai C. *Mining Text Data*. New York, NY: Springer (2019).
5. Liu B. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge: Cambridge University Press (2020).
6. Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press (2018).
7. Nguyen TT, Aberer K. Clustering and classification of social media data. *Soc Net Anal Mining.* (2018) 8(1):1–14.
8. Zhang Y, Wang J, Zhao X. An improved fuzzy clustering algorithm for large-scale text data analysis. *Expert Syst Appl.* (2020) 158:113–21.
9. Blei DM. Probabilistic topic models. *Commun ACM.* (2012) 55(4):77–84.