# Analyzing Social Media with an Improved K-Means Clustering Algorithm

Sujeet Kumar Sahani[1] and Ms.Sonam Singh[2]
[1]*Research Scholar, SHEAT College Babatpur, Varanasi*
[2]*Assistant Professor, SHEAT College Babatpur, Varanasi*

**Abstract- Nowadays, sharing human social behavior and growing a multi-user network requires the use of social media. Analysis of social media provides a valuable opportunity to study human social activity at scale. People often bring and talk about different kind of topics on social media platforms. depending on which, discussions are made to show the positive as well dis positives ways. This is an interesting point about using social media. There is a great deal of information about the type and substance of these discussion holography's that informs us more broadly around patterns in social media interactions, and how content flows between individuals.**

**Data mining techniques exists which can be utilized to crunch out user information from social media and relationships existing within the network. However, state of the art methods often fail to model user communities and their behaviors at a correct level. To solve this problem and help successfully do the grouping of related information, in our work here we present an enhanced fuzzy means clustering approach. Using this technique reduces the creation and time complexity of clusters, resulting in higher quality group outcomes.**

**Our proposed system aims at mitigating the problems as described above by introducing a scalable and effective solution for real-time cluster social media data.**

**Keywords: Social networking, clustering, Java programming, Hadoop, K-means, Big data**

## 1. INTRODUCTION

Clustering is the process of dividing a collection patterns into separate groups or clusters. This approach ensures that patterns in different clusters are different from one another, and more similar to the other dynamics within their same cluster. Many discipline like artificial intelligence, statistics and neural networks have been researched extensively in field of clustering. Helps finding patterns in data and organizing similar items into one group. Clustering The basic concept of clustering is intuitive, and mirrors how humans break the huge amounts of data we have into more manageable subsets.

However, manual labeling of data is a big challenge especially in high-dimensional (more than 2 or 3 dimensions) case. Several methods have been developed to solve this problem, typically referred as "Data Clustering Methods" which are soft computing techniques. The models that can be constructed using these techniques are also used for data compression on top of organizing and sequestering the data into classes. We describe the similar kind of data using less number of symbols and create some models by checking what are common in this group wise.

To cluster data effectively involves the existence of natural groupings within a particular dataset, otherwise clustering effort can be vain and lead to random partitions. Another challenge of clustering is the possibility that there may be overlap between data groups, which reduces the effectiveness of this method. This could be addressed in a future study as other clustering version methods may be applied to both fuzzy or more sophisticated neural networks. Second, they are commonly employed as preprocessing schemes to derive initializations for fuzzy if-then rules or radial basis functions.

The purpose of every clustering methods discussed here is to find the cluster centers that representing each clusters. It works by comparing a new input vector to the cluster centers, then categorizing it with which ever group its feature values are most similar too. Some clustering approaches, such as Fuzzy C-means and K-means clustering require knowing in advance the number of clusters. The procedures which are followed in here, help to divide the data into the

number of clusters specified. However, the above methods are only effective when number of clusters is known in advanced.

Clustering algorithms can be of two types: Offline and Online. Online clustering continuously updates cluster centers with every new input vector to provide real-time learning. Offline clustering on the other hand discovers cluster centers using a set of training data, fixes them and uses an incoming input vector for categorization by finding which its closest matched center is. All the techniques discussed in this paper are offline based.

In the following sections, we will discuss in detail four different clustering algorithms.

EXISTING SYSTEM:

K-means is one of the earlier algorithms existing for this purpose, although has some significant drawbacks. Problem: of One the biggest problem with K-means is that it often leads to bad clustering and also, K means can easily fall into local minimum because when we reduce sum of squared distance in each iterations which doesn't ensure global optimum. Also, K-means has to deal with the problem of centroid initialization but it faces challenges when dealing with Density-Based spatial clustering (which in turn could very well lead poor cluster formation). To enhance the effectiveness and efficiency of cluster, we recommend to use bisecting optimal cluster distance algorithm apply.

PROPOSED SYSTEM:

The amount of information circulating social media has increased exorbitantly over the past few years. Retailers and financial institutions can greatly benefit from an analysis of social media data when it comes to understanding user behavior. Retail companies are seen questioning about risk-prevention strategies, customer Relationship Management and branding through social media. The correlation between market behavior and Twitter data, as a proxy for user sentiment is just one of the examples.

While need for Social Network Analysis with sociological underpinning has exploded alongwith the newsworthy blogs and discussion forums. Huge databases, partly formulated from user involvement in content generation on social media sites have forced the development of advanced data mining algorithms. For instance in social media, the news forum Slashdot can help access more detailed user information from sites like Facebook for a cleaner clustering method. The framework employs an optimized fuzzy means clustering algorithm that is more precise than existing methods.

Clustering is a technique used to classify things into groups in exploratory data analysis where there is high similarity between members of those groups. Unsupervised learning- here we divide data into groups of similar objects, or dissimilar. According to the authority model, our approach needs to differentiate between user groups on grounds of behavior and attitudes.

For each example, the standard K-means algorithm has to calculate a distance from every cluster center this takes time. Procedure of K-means algorithm :

1. Load the twitter dataset on server side.
2. Randomly select centroids from the dataset, insert number of clusters
3. Calculate how far these data points lie from their respective centroids using Manhattan distance.
4. For each data point, assign it to the cluster whose centre is closest.
5. Repeat steps 3 and 4 until centroids stop fluctuating.
6. Retrieve the clustered data stored in -clustered. Done
7. Calculate the SSE of each cluster.

Eliminating time and space complexity of the K-means algorithm,as the solution to this problem, herein is proposed an optimized fuzzy-means- clustering algorithm as:

Optimization of Fuzzy Means Clustering Algorithm:

Input:

- Number of Clusters, k
- Dataset D = (d1, d2, …….dn) n data objects

Output:

- A set of k clusters

Steps:

1. Initially selectk data points randomly as the first centers Cluster 1 Clustering ofdistribute D.
2. Find the matched terms for each cluster center cj ($1 \leq j \leq k$) and every data object di ($1 \leq i \leq n$).
3. Calculate the sum of squared errors (SSE).

4. Calculate how much each data point depends on the clan by calculating exactly many words were in both the datapoint and its centroid
5. By weight, determine which data points are closest to the centroids.
6. After writing the weight of data points for each cluster center cj ($1 \leq j < k$), you assign data items di to the closest cluster.
7. Recomputed cluster centers.
8. Repeat above process until the kernels of clusters stop.
9. Display clusters result

By this advanced method the drawbacks of traditional k-means algorithm can be overcome which in turn improves the accuracy and productivity of social media data grouping.

Operational Feasibility:

The utility of proposed measures is determined by its possible integration into an information system that meets the operational requirements of the organization. In order to provide the implementation, every project must have some operability. Factors to Consider for Assessing Operational Viability

1. Are users and management supportive?
2. How will the system be constructed and installed, how it can function productively?
3. Are there any user means of reducing these potential pitfalls to the point where that negates the application's benefits?

The design of this system rectifies these matters. This ensures it provide useful benefits that users can utilize and addresses concerns of management, again ensuring that there no user resistance to work with this application. Also, this planned design will ensure better efficiency of your computer resources and hence overall performance.

## 2. SYSTEM DESIGN

UML Diagrams:

The final part of the system design is to identify what architecture, parts, modules and interfaces are required in order for a request (determined by previously stated criteria) to be satisfied. It is really product development being applied to systems theory. It combines systems engineering, with systems architecture and analysis in complex integrated system development. If product development brings design, manufacturing and marketing into a coherent way of thinking on how to make customers happy with the best possible abilities at last making something that will become what heuristics seek only recently in examples: marketed drivel.

System Design - Develop and Construct the Systems that Meet User Needs. System design was a critical and often appreciated role in the data processing part of IT until probably sometime during the 1990s. However, hardware and software characteristic of 1990s has allowed to build modular systems. Of course the importance of software engineering has grown with it, as soon as software running on (more or less) generic systems started to play a big role.

Object-oriented analysis and design methodologies have enjoyed considerable usage in the design of computer systems. Unified Modeling Language (UML) The de-facto industry standard language for object-oriented analysis and design It is most used to model software systems but increasingly being applied more broadly for organizational and non-software structures.

Use Case Relationships:

There are three correlations between use cases that are frequently used:

Elements of a Collaboration Diagram:

The components of a collaboration diagram are as follows:

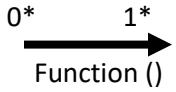
Object:It refers to the entities that participate in systemic interactions. Here they are represented as a rectangle with the name of what is below object in front and again followed by colon.

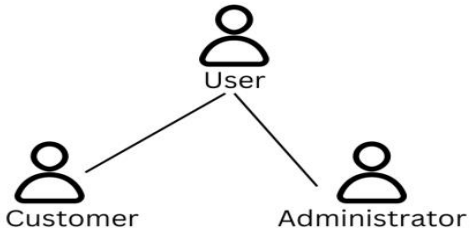

Relation/Association: This is the link that binds together related items. for show cardinality, qualifiers can be put earlier than both finish of the connection.



\

Messages:To show that the initiating object is interacting with target objects, an arrow represents from them. The arrows (with numbers) indicate sequence of these interactions process.






3. LANGUAGE SPECIFICATIONS

Java Technology

Java Architectural Framework stands on 4 linked technologies.

1. The Java programming language
2. The Java class file format
3. Java API (Application Programming Interface )
4. Java Virtual Machine (JVM)

While we write and execute a Java program these four technologies are in-place. This is how it operates:

In the Java programming language, you create source files and then compile them.

These source files are then compiled to Java class-files

Running the class files (JVM) with Java virtual machine

While creating your program, you call system resources such as I/O activities using methods from the Java API. Your software fulfills these API calls, running the class files that implement them when it runs. This gives us a clear picture of the way in which each part plays its role and interacts with other parts of Java design.

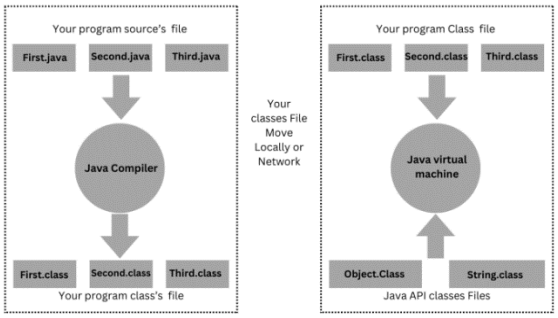



Fig. Java Programming Environment

Systems Development Life Cycle

Systems in Software engineering/Information Systems/System Engineering could be developed through a lifecycle called System Development Life Cycle (SDLC). It models a family of systems and the different types of development approaches they can use. A set of steps or phases of development and maintenance is the SDLC or software life cycle which offers structure in organizing and managing all type our projects can be considered as software, even planning to developing testing etc., that's much important where each methods we use need for their work.

What is SDLC ?

Keeping aside that SDLC (Software Development Lifecycle) is a term given to the processes involved in producing software; starting from planning through implementation, testing and deployment. With the tools and utilities at their disposal, software developers complete a series of tasks using various methods according to particular project requirements known as SDLC(software development life cycle) models. Prescriptive: how software development should ideally go Descriptive: how a specific software system was produced Both prescriptive and descriptive models are used to build predictive ones Prescriptive models provide structured guidelines for software development, and descriptive models help understand the process of developing a software.

SDLC Models

SDLC models are of different types, as given subsequently: -

1. Linear model (Waterfall) : The process works linearly, each of the phases is distinct and completed in turns like development & specification etc. This is purely sequential, you are not allowed do the next step until all in previous (higher) level has been finished.
2. Integrated Evolutionary architecture: Development and specification are performed simultaneously in this type of architectural pattern. As examples, consider:
3. Incremental Model: waterfall approach with incremental development adding phase to the SDLC by more focus on testing process thus decreasing risk.
4. RAD (Rapid Application Development): It is an object-oriented technology, which suits best to develop high-quality products quickly.

5. Spiral Model: This involves component development using a process starting with tiny modules that spiral out.
6. Formal systems development : In this method, an implementation is created using a mathematical model of the system.
7. Agile Methods: These methods handle the process with a lighter touch and focus on being flexible at quickly responding to needs or input changes, such as those seen in RUP.
8. The reuse-based development : a system is built by assembling existing parts and reusing proven solutions.

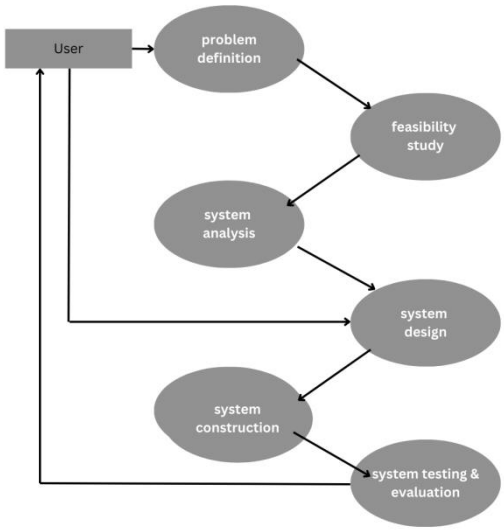



Fig. Basic Step of SDLC

Spiral Life Cycle Model:
The spiral model, like incremental is also one of the development models where risk analysis plays vital role. It operates by means of repetitive cycles or "spirals," with the four major stages being planning, risk assessment, engineering and evaluation.

Here's how it works:

- Planning Phase: Risks are identified and Initial requirements are gathered.
- Risk Analysis Phase: Risks should be assessed, and alternatives are proposed. Maybe you could prototype some of these ideas and test them.
- Engineering Phase: Software is developed, tested and refined.
- Evaluation: Before next spiral is identified client will evaluate the iteration results and progress on a large scale.

In the spiral model, radius stands for related costs and angular movement around the spiral represents progress. This iterative approach allows for flexibility and continuous improvement over the lifespan of a project.

## 4. TESTING

Testing Methods
The Box Approach
Software testing techniques are typically divided into two main categories: white-box and black box-testing. These are the different categories of test cases creation as it depends on various approaches, which a Test Engineer may consider.

White Box Testing
WHITE-BOX TESTING: This testing provides the tester with real codes, data structures and algorithms. It allows to perform complete and detail-oriented testing. That is White-Box Testing Types.

API Testing: This part includes testing the application with both public APIs, as well as secret ones

Code Coverage: Creating tests that require every segment of a code to be executed at least once.

Fault Injection Strategies: Adding faults to the code increases test coverage and stress tests multiple paths.

Mutation Testing Techniques : changing a few lines of code and checking whether the tests will identify issues.

Static testing: Testing of the code, looking for defects with actually executing it.

Black Box Testing
Black box testing treats the software program as a "black box", i.e., does not take into account its internal structure. The whole point would be to test outside behavior. Black box testing techniques include.

Equivalence Partitioning : classifying the input data in various parts that have predicted similar outcomes.

Boundary Value Analysis : testing along the partitions.
All-Pairs Testing: Consideration each conceivable pair of the parameters.

Fuzz Testing: inject random information to expose vulnerabilities and devise unusual behavior.

Model-Based Testing: Where a model is used to generate test cases.

Traceability Matrix: making sure that each and every requirement has been checked

Exploratory Testing: Learning, Creating & even Implementing assessments all at the same time.

Specification-Based Testing: testing the program based on its specs and requirements.
In the case of specification-based testing, test data is entered and the output checked against what was expected This technique ensures that the software doing what it is suppose to do with respect of its requirements. It may not, however, find every potential issue so it is important but insufficient by itself.

## 5. CONCLUSION

In this paper, we have proposed a novel and simple clustering algorithm that requires much less space complexity and time complexity than other complex algorithms. More exactly we have showed a better fuzzy means clustering algorithm which significantly refines the dataset's clusters. It is a very simple concept: how much does some data object weighs, with respect to the centroids. This weight computation is required far fewer times as compared to the standard k-means approach which further repeats until centroids stable. This reduces the runtime because each datum converges to its cluster center faster. This suggested solution implementation speeds up the clustering process and improves its accuracy by lowering computing complexity of conventional k-means.

## REFERENCE


[1] A. Ghosh and T. Veale (2016) ‘‘Fracking sarcasm using neural network,’’ in Proceedings of 7th Workshop Computing Approaches Subjectivity, Sentiment Social Media Analysis, pp. 161–169.

[2] A. Joshi, P. Bhattacharyya, and M. J. Carman (2017)‘‘Automatic sarcasm detection: A survey.’’ ACM Computing Surveys, Article Id:1000, CSUR, ACM Digital Library, pp.1-22.

[3] A. Joshi, P. Bhattacharyya, M. Carman, J. Saraswati, and R. Shukla (2016) ‘‘How do cultural differences impact the quality of sarcasm annotation?: A case study of indian annotators and american text,’’ in Proceedings of 10th SIGHUM Workshop Language Technology Cultural Heritage, Social Science and Humanities., Association of Computational Linguistics, Berlin, Germany, pp. 95–99.

[4] A. Joshi, V. Sharma, and P. Bhattacharyya (July 2015) ‘‘Harnessing context incongruity for sarcasm detection,’’ in Proceedings of 53rd Annual Meeting Association of Computational Linguistics, International Joint Conference Natural Language Process. (ACL-IJCNLP), vol. 2, pp. 757–762.

[5] A. Kumar and A. Jaiswal (October 2017) ‘‘Empirical study of X.com and Tumblr for sentiment analysis using soft computing techniques,’’ in Proceedings ofWorld Congress on Engineering and Computer Science (WCECS 2017), San Fransisco, USA,pp. 1–5.

[6] A. Rajadesingan, R. Zafarani, and H. Liu (February 2015) ‘‘Sarcasm detection on X.com: A behavioural modeling approach,’’ in Proceedings of 18th ACM International Conference on Web Search and Data Mining (WSDM), Shanghai, China, pp. 79–106

[7] A. Reyes, P. Rosso, and D. Buscaldi (2012) ‘‘From humor recognition to irony detection: The figurative language of social media,’’ Data Knowledge and Engineering, vol. 74, pp. 1–12.

[8] A. Reyes, P. Rosso, and T. Veale (2013) ‘‘A multidimensional approach for detecting irony in X.com,’’ Language Resources Evaluation., vol. 47, no. 1, pp. 239–268.

[9] A. Utsumi (September 1996)“Implicit display theory of verbal irony: Towards a computational model of irony," in Proceedings of the 12th -20 workshop on language technology joint with International Workshop of Computational Humor, Amsterdam.

[10] Abhijit Mishra, DipteshKanojia, and Pushpak Bhattacharyya (2016)“Predicting readers sarcasm understandability by modeling gaze behavior”, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), Arizona, USA.