IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SOCIAL MEDIA ANALYSIS USING CLUSTERING ALGORITHM

[1]Mrs. Aarti S. Jadhav-Patil, [2]Sharli Salokhe, [3]Pooja Duddalwar, [4]Nikita Kakade, [5]Akanksha Chavan

[1]Professor, [2]Student, [3]Student, [4]Student, [5]Student
[1]Department of Information Technology,
[1]D. Y. Patil College of Engineering, Akurdi, Pune, India

Abstract: Social media is a platform which uses its popularity to discuss various problems whether it is related to political views, Healthcare, Finance or from other domains. In the 21st century Social media has a huge impact on people's lifestyle & their perspective. In this project, We tend to create groups & participate in various discussions through the social media community. For example, users who are social workers and introduce new issues and discussions on social media. Furthermore, positive or negative attitudes can also be inferred from those discussions. Such problems require a formal analysis of social media logs and units of information that can spread from person to person through the social network. Analytics persons and businesses feel the need to gain new insights from social media; they require the analytics tools and expertise to transform this information which will have a big volume and variety into the respective strategies to draw certain results. Social media analytics is a useful tool for getting details of customers that are distributed across online sources. Social media has acquired immense popularity and interest in marketing teams.

***Key Words* – Clustering Algorithm, Social media, Machine Learning**

1. Introduction

Social media serves as a ubiquitous public platform. Within the applications, the user creates individual unique expressions for data exchange. Social Media Analysis means checking the taste, views, and interests of people regarding celebrities, politicians, or some other topic. The basic thing Social Media Analysis does is to classify opinions into different categories like positive, negative, and neutral. For example, Analysis of a celebrity, how do people think about him? Whether they think positively, or negatively somewhere in between them. On Twitter, tweets of different users produce a larger amount of data and share the information in the form of an unstructured type of data. It is difficult to understand and extract information from data. Hence, we need tools and technologies that can store and process unstructured and big data. There are different techniques and tools available that can handle this type of data, we have used the K-Means, K- Medio's, and Agglomerative Hierarchical clustering algorithms. Social media is a web based and mobile-based internet application that will allow the formation, access, and exchange of user generated content that is universally accessible. Besides social networking media like Twitter and Facebook, the term social media encompasses really simple syndication (RSS) feeds, blogs, wikis, and news, typically yielding unstructured text accessible through the web. Social media is important for research into computational social science that use to explore questions using quantitative techniques. Social media has led to numerous data services, gears, and analytics platforms. The tools available to researchers have either given superficial access to the raw data or non-superficial access. Researchers necessitate programming analytics in a language such as Java. So the proposed work is much better than the available ones concerning cost, efficiency, and scalability The analytics persons and businesses feel the need to gain new business insights from social media; they require the analytics tools and expertise to transform this information which will have big volume and variety into the respective strategies to certain result.

The detection of the social network structure is an important analysis area of social media analysis. Detecting communities is a major focus in sociology, computer science, biology, and methods, where systems are typically represented as graphs. The goal of community detection is to find clusters as subgraphs inside a given network. With the internet's democratization, communicating and sharing information is easier than ever. Community detection is a method for comprehending the structure of complicated networks and, ultimately, collecting meaningful information from them.

2. LITERATURE REVIEW

In the past few years, several clustering algorithms for big data have been proposed which are derived based on distributed and parallel computation. In 1967 Mac Queen was the first to propose this technique. The first standard algorithm was proposed by Stuart Lloyd in 1957 as an approach to pulse-code modulation. Oftentimes it is referred to as Lloyd-Fogy, because in 1965, E.W. Forgy published essentially the same method. According to K.A. Abdul Nazeer, all the major drawback of the k-means algorithm is about choosing the initial centroids which produce different clusters. But eventually, the final cluster quality of algorithms depends on the selection of initial centroids, chosen at the time of computation. According to Y. S. Thakare et al., the performance the of K-means algorithm is evaluated with various databases such as Iris, Wine, Vowel, Ionosphere, and Crude oil data Sets and various distance metrics. It is concluded that the performance of k-means clustering is dependent on the database used as well as distance metrics. Soumi Ghosh et al. proposed a comparative discussion of two clustering algorithms namely centroid-based K-Means and representative object-based Fuzzy C-Means clustering algorithms. This discussion is based on the performance evaluation of the efficiency of clustering output by applying these algorithms. The result of this comparative study is that FCM produces a closer result to the K-means but still, computation time is more than the k-means due to the involvement of the fuzzy measure calculations. Sakthi et al. proposed that due to the increment in the amount of data across the world, analysis of the data turns out to be a very difficult task. To understand and learn the data, classify those data into remarkable collections. So, there is a lack of data mining techniques. Amutha et al. proposed that when two or more algorithms of the same category of clustering technique are used the best results will be meant. Two k-means algorithms: Parallel k/h-Means Clustering for Large Data Sets and A Novel K-Means Based Clustering Algorithm or High Dimensional Data Sets. The parallel k/h-Means algorithm is aimed to deal with very large data sets. Novel K-Means Based Clustering provides the advantages of using both HCand K-Means. Using these two algorithms, space and similarity between the data sets present in each node are extended. Nidhi Singh et al. proposed the comparative analysis of one.

The problem of the people living in a particular area detection has been widely studied within the context of large-scale Community detection algorithms attempting to identify groups of vertices more densely connected than the rest of the network. Social network extracts from social media however present unique challenges due to their size and high clustering coefficients.

The problem of community recognition has been widely studied within the context of large scale. In the existing technique, performing clustering of social media information faced so many problems .Social networks present unique challenges due to their size and high clustering coefficients.

3.1 PROPOSED SYSTEM

Clustering structures the data into a collection of objects that are similar or dissimilar and is considered unsupervised learning. The application method is mainly on finding user groups based on activities and attitude features as suggested in the authority model. Social media serves as a ubiquitous public platform. Within the applications, the user creates individual and unique expressions for data exchange. The study of this data by industry specialists seeking new and inventive methods to collect data for analysis remains important to the future of social media.

3.2 SYSTEM ARCHITECTURE

In data mining, two learning styles are used to mine data i.e., supervised learning and unsupervised learning.

Supervised learning in this learning, data includes together the input and the asked result. Unsupervised learning the asked result isn't handed to the unsupervised model during the learning procedure. This system can be used to cluster the input data.

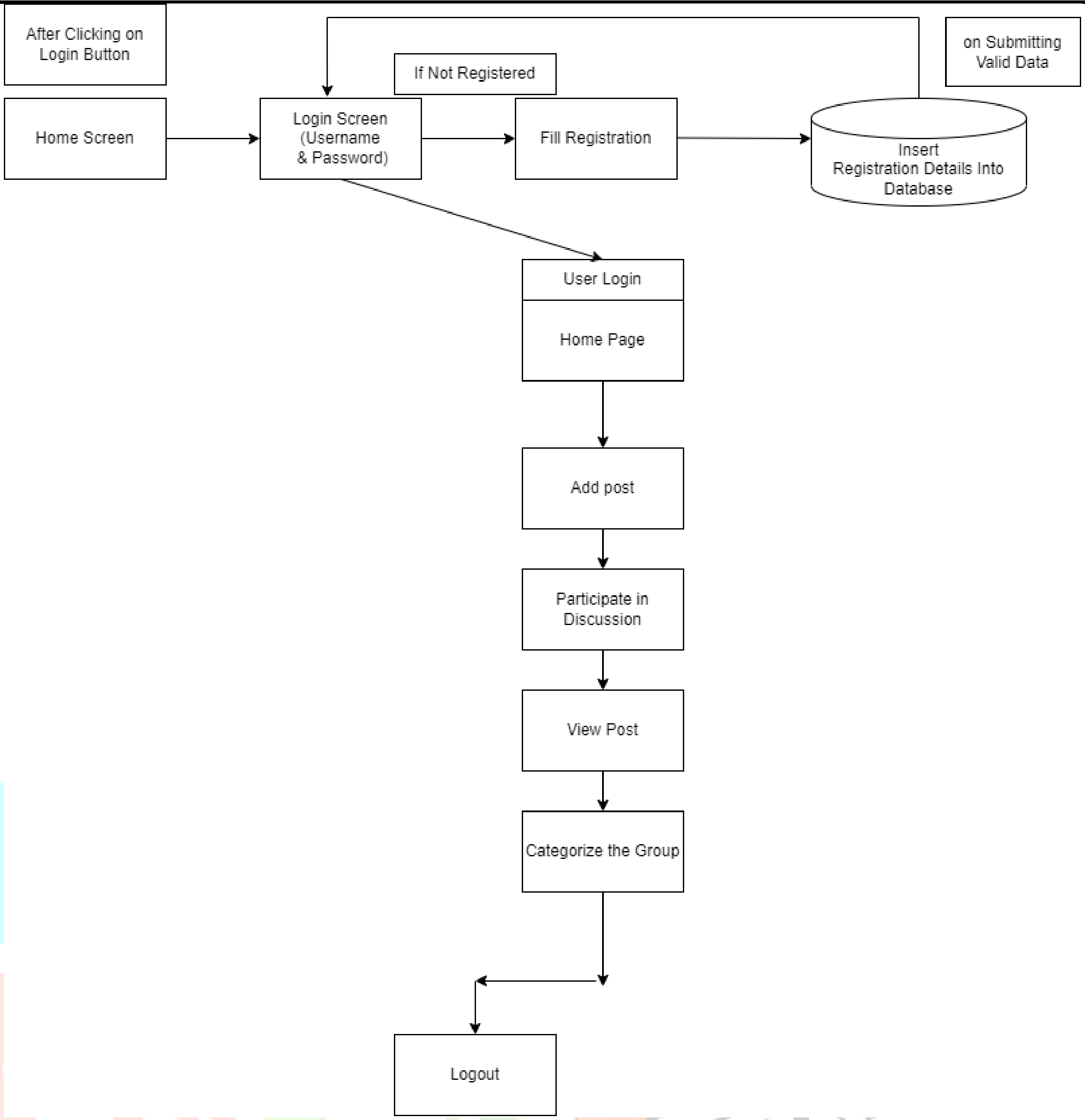

Fig 3.2.1

3.3 Module Description

3.3.1 Building the Network:

Generate a social network and ameliorate the quality and representation of the operating graphs. We will filter some of the comments according to the following four criteria. The post-Anonymous comments were also discarded. We discard very low-quality comments with the score.

3.3.2 Degree Distributions:

The analysis of this degree distribution describes the level of interaction between users and provides a robust indicator of the grade of heterogeneity in the network.

3.3.3. Community Structure:

The k-Means algorithm is used to cluster data and detect communities by clustering messages from the large stream of social data. This module helps to identify a group of people involved in the discussion.

4. SOFTWARE REQUIREMENTS

4.1. Python

- Python is a powerful multi-purpose programming language. It has simple syntax, making it the perfect language for someone trying to learn computer programming for the initiatory.
- This is a broad scope on how to get started in Python, why you should learn it and how you can learn it. However, if you know other programming languages and want to quickly get started with Python.
- Python is a general-purpose language. It has a wide range of applications from Web development like Django and Bottle, scientific and mathematical computing to desktop graphical user Interfaces like Pygame, Panda3D. The syntax of the language is clean and the length of the code is comparatively short. It's fun to work in Python because it allows you to think about the problem rather than aim on the syntax.

4.2. Django

- Django is a high-level Python Web framework that stirs up rapid development and clean, pragmatic design. Built by knowledgeable developers, it takes care of much of the inconvenience of Web development, so you can focus on writing your app without needing to reinvent the wheel. It’s free and open source.
- Features of Django are:
    - Rapid Development
    - Secure
    - Scalable
    - Fully loaded
    - Versatile
    - Open Source
    - Vast and Supported Community

4.3. XAMPP Server

- XAMPP is a Windows OS-based program which is designed to offer an easy way to install Apache, PHP and MySQL packages with an easy-to-use installation program instead of having to install and configure everything yourself.
- XAMPP is so easy because once it is installed it is prepared. You don’t have to do any additional configuration of any configuration files to get it running.

4.4. PHP

- Allows you to change or add users and for making new databases phpMyAdmin is a free software tool written in PHP, intended to handle the administration of MySQL over the World Wide Web. phpMyAdmin supports an extensive range of operations with MySQL.
- The most frequently used operations are supported by the user interface like managing databases, tables, fields, relations, indexes, users, permissions, etc., while you still can directly execute any SQL statement.

4.5. MySQL

- SQL Server is a relational database management system from Microsoft which is designed for the endeavour environment.
- SQL Server runs on Transact SQL, a set of programming extensions from Sybase and Microsoft that add several features to standard SQL, including transaction control, exception and error handling, row processing, and declared variables.

4.6. CLUSTERING

4.6.1. Overview

Clustering is the method of partitioning the population or data points into a number of groups such that those data points in the same groups are more similar to other data points in the same group than those in the other groups. In simple words, the aim is to segregate groups with specifically similar traits and assign them into clusters. Let’s understand clustering with an example. Let’s say, you are the head of a rental store and want to understand the preferences of your customers to scale up and scale out your business. Is it possible and feasible for you to look at the details of each customer and devise a unique business strategy for each and every one of them? Definitely not. But, what can be done is to cluster all of your customers into groups based on their purchasing habits and spending habits and use a separate strategy for customers in each of these 10 groups. And this is called a method of clustering.

4.6.2. Types of Clustering

Statistically speaking, the methods of clustering are divided into two subcategories Hard Clustering: In hard clustering, every data point either belongs to a cluster completely or not. For example, in the above given example every customer is put into one group out of the total number of groups. Soft Clustering: In soft clustering, in place of putting each data point in a separate cluster, a probability or likelihood of that data point to be in these clusters is assigned. For example, from the above scenario each customer is assigned a probability to be in either of 10 clusters of the retail store.

4.6.3. Types of clustering models

Since the result of clustering is subjective, this means that there are multiple possible methods that can be used for achieving this goal. Every methodology follows a different set of rules for defining the 'similarity' between data points.

A. Connectivity Models

As the name suggests, these models are based on the idea that the data points closer in data space show more similarity to each other than the data points situated farther away. These models follow two following approaches. In the first approach, they start with classifying all data points into separate clusters & then aggregating them as the distance between these data points decreases. In the second approach, all data points are categorized as a single cluster and then partitioned as the distance between the data points increases. Also, the choice of distance function is totally subjective. These models are very easy to interpret but lack the scalability for handling huge datasets. Examples of these models are hierarchical clustering algorithms and its forms.

B. Centroid Models

These are iterative clustering algorithms in which the idea of similarity is derived by the closeness of a data point to the centroid of the clusters. K-Means clustering algorithm is a popular algorithm that falls into this classification. In these models, the number of clusters required at the end have to be mentioned beforehand, which makes it important to have prior knowledge of the dataset. These models run iteratively to locate the local optima.

C. Distribution Models

These clustering models are based on the idea of how probable it is that all the data points in the cluster belong to the same distribution (For example: Normal or Gaussian). These models often suffer from overfitting. A popular example of these models is the Expectation-maximization (EM) algorithm which makes use of multivariate normal distributions. Density Models

These models search the data space for areas of different density of data points in the given data space. It separates various different density regions and assigns the data points within these regions in the same cluster. Popular examples of density models are DBSCAN and OPTICS.

5. HARDWARE REQUIREMENT

i. Laptop or PC
ii. Windows 7 or higher
iii. i3 processor system or higher
iv. 4 GB RAM or higher
v. 100 GB ROM or higher

6. CONCLUSION

This is the project of System Design about "**The Social Media Community Using Optimized Clustering Algorithm"** developed in Django in Python programming language. This project proposes different things to analyse social media data. In this method, the method used to fetch data is a face pager. The fetched data will be exported into CSV files. By separating the post comments according to the activities and finding results and making the communities based on the result of the post comment. To create a framework for the unique and unexpected task of detecting communities by clustering messages from large streams of social data. Our framework uses the K-Means clustering algorithm along with the Genetic algorithm and Optimized Cluster Distance method to cluster data.

7. REFERENCES

[1] Refining Initial Points for K-Means Clustering Paul S. Bradley Usama M. Fayyad, Appears in Proceedings of the 15th International Conference on Machine Learning (ICML98), J. Shavlik (ed.), pp. 91-99. Morgan Kaufmann, San Francisco, 1998.
[2] A Comparative Study of K-means and K-medoid Clustering for Social Media Text Mining Volume 2 ||Issue 11 ||JUNE 2017||ISSN.
[3] Analysis and Visualization of Twitter Data using k-means Clustering, International Conference on Intelligent Computing and Control Systems ICICCS 2017.
[4] Kalra, M., Lal, N., & Qamar, S. (2017). K-Mean Clustering Algorithm Approach for Data Mining of Heterogeneous Data.
[5]www.Geeks for Geeks .com