HEALTH MEDICARE DATA USING TWEETS IN TWITTER

DHIVYA .S [1], NITHESH KUMAR .G [2], SHIDA SABARI .K [3]

[1]*Assistant Professor, Department of CSE, Jeppiaar SRR Engineering College, Padur, Chennai.*
[2, 3] *Final Year Student, Department of CSE, Jeppiaar SRR Engineering College, Padur, Chennai.*

Abstract - *Web-based social networking furnishes boundless chances to impart encounters to their best recommendation. In current situations and with accessible new advances, twitter can be utilized adequately to gather data as opposed to social affair data in conventional technique. Twitter is a most prevalent online long range informal communication benefit that empowers people to share and pick up information. Here we analyze that which country/place having frequent talks about particular disease. We can predict where the effect of the disease will high. This framework manages the difficulties that show up during the time spent Sentiment Analysis; continuous tweets are considered as they are rich wellsprings of information for assessment mining and feeling examination. The fundamental goal of this framework is to perform constant nostalgic examination on the tweets that are extricated from the twitter.*

Key Words: **Twitter, SVM, Classification, Similarity Clustering, HDFS.**

1. INTRODUCTION

Facts mining is an interdisciplinary subfield of laptop science. it is the computational technique of discovering patterns in big statistics units concerning methods at the intersection of artificial intelligence, device gaining knowledge of, information, and database structures .In recent times, social network structures are the getting popular in which hundreds of thousands of users can give their views approximately any product. Sentiment evaluation gives a powerful and green approach to show public opinion well timed which gives critical records for decision making in diverse domain names. For obtaining customers feedback closer to any product, one of a kind agencies can take a look at the general public sentiment in tweets. Many research and business applications have been done within the location of public sentiment monitoring and modeling. It's been suggested that events in actual lifestyles certainly have a significant and instant effect on the public sentiment in on-line. But, none of these research finished further analysis to mine beneficial insights in the back of significant sentiment version, known as public sentiment version.

2. PROPOSED SYSTEM

The proposed framework utilizes Twitter to get the data and process on it. The data from the Twitter is removed utilizing crawing and Twitter API. The twitter API will slither the tweets from twitter utilizing twitter4j. Twitter4j will separate the tweets and show to the client I the table arrangement. These separated tweets are then preprocessed by supplanting the short frame words with full shape. It likewise expel the stop words frame the separated tweets are then put away in the database. The pre-processed tweets are additionally arranged utilizing SVM grouping in light of the classification

3. SYSTEM IMPLEMENTATION

Cutting-edge large facts technologies make it possible in a short time to examine a huge collection of information from heaps of patients, pick out clusters and correlations, and increase predictive fashions the usage of statistical or machine-mastering modeling techniques. a good way to examine complicated statistics and to pick out styles it is very crucial to soundly store, control and proportion large quantities of complicated records. Cloud comes with an express protection assignment, i.e. the records owner may not have any control of where the statistics is located. Hadoop, it is less complicated for businesses to get a grip on the massive volumes of information being generated each day, but at the identical time can also create troubles related to protection, information get entry to, monitoring, excessive availability and commercial enterprise continuity. Hadoop has two fundamental sub tasks – Map lessen and Hadoop distributed report gadget (HDFS). MapReduce is a framework for processing parallelizable problems throughout massive datasets using a huge variety of computer systems (nodes), collectively called a cluster (if all nodes are at the identical nearby community and use comparable hardware) or a grid (if the nodes are shared across geographically and administratively disbursed structures, and use extra heterogeneous hardware).

Processing can arise on information saved either in a record device (unstructured) or in a database (dependent). MapReduce can take benefit of locality of information, processing it on or near the storage assets for you to lessen the space over which it should be transmitted. The core of Apache Hadoop consists of a storage component (Hadoop disbursed report system &#40;HDFS&#41;) and a processing part (MapReduce). Hadoop splits documents into large blocks and distributes them among the nodes in the cluster. To procedure the statistics, Hadoop MapReduce transfers packaged code for nodes to technique in parallel, primarily based on the facts each node wishes to system. The Hadoop disbursed report gadget (HDFS)—a subproject of the Apache Hadoop mission—is a allotted, especially fault-tolerant record device designed to run on low-value commodity hardware. HDFS provides high throughput get entry to software records and is appropriate for packages with large facts units.

The approaching venture in healthcare is "operating with big data in clinic systems is hugely challenging however at the same time holds super promise in presenting greater significant records to assist clinicians treat sufferers throughout the continuum of care". private health file &#40;PHR&#41; is an emerging patient-centric model of fitness data alternate, that's often outsourced to be saved at a third birthday party, inclusive of cloud vendors. However, there were extensive privateers concerns as personal fitness records might be exposed to the ones third celebration servers and to unauthorized events.

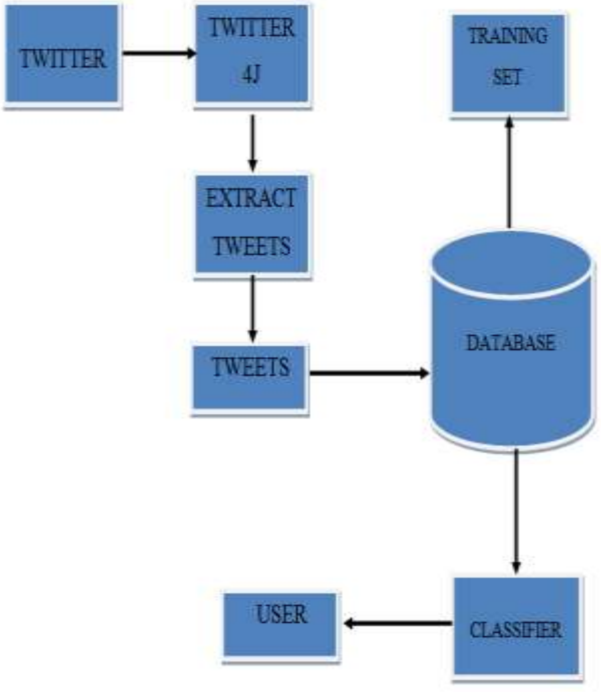

Figure 1: System Design Architecture

Implementation is that the stage of the project when the theoretical design is became a working system. Thus it are often considered to be the foremost critical stage in achieving a successful new system and in giving the user confidence that the new system will work and be effective. There are various Features and Benefits in implementing the java language in the project.

3.1 Platform Independent

The concept of Write-once-run-anywhere (known because the Platform independent) is one of the important key feature of java language that makes java because the foremost powerful language. Not even one language is idle to the present feature but java is closer to the present feature

3.2 HDFS (Hadoop Distributed File System)

The Hadoop disseminated record framework is the essential information stockpiling framework utilized by Hadoop applications. It utilizes name hub and information note design to actualize a circulated record framework that gives superior access to information across exceptionally versatile Hadoop groups. The center of Apache Hadoop comprise is of a capacity part and a preparing part. Hadoop parts documents into enormous squares and circulates them among the hubs in the group. To process the information, Hadoop Map Reduce moves bundled code for hubs to process in equal, in view of the information every hub needs to process. The Hadoop is a subproject of Apache Hadoop venture which is a circulated, profoundly shortcoming tolerant record framework intended to run on minimal effort ware equipment. HDFS gives throughput access to application information and is reasonable for application with enormous datasets

3.3 USE CASE DIAGRAM

In software and systems engineering, a use case could also be an inventory of steps, typically defining interactions between a task (known in UML as an "actor") and a system, to understand a goal. The actor are often a person's or an external system. In systems engineering, use cases are used at a better level than within software engineering, often representing missions or stakeholder goals. The detailed requirements may then be captured in SysML or as contractual statements.

3.4 CLASS DIAGRAM

The class diagram shows how the various entities (people, things, and data) relate to every other; in other words, it shows the static structures of the system. A class diagram are often wont to display logical classes. Class diagrams also can be wont to show implementation classes, which are the items that programmers typically affect . A class is depicted on the class diagram as a rectangle with three horizontal sections, as shown in above figure. The upper section shows the class's name, the center section contains the class's attributes, and therefore the lower section contains the class's operations (or "methods"). The diagram has five main classes which give the attributes and operations used in each class.

3.4 COLLABORATION DIAGRAM

The concept is quite a decade old although it's been refined as modelling paradigms have evolved. These labels are preceded by colons and may be underlined.

3.5 SEQUENCE DIAGRAM

A sequence diagram during a Unified Modeling Language (UML) may be a quite interaction diagram that shows how processes operate with each other and in what order. Sequence diagrams typically are related to use case realizations within the Logical View of the system under development.

4. MODULES

The System module is categorized into four sub-modules namely,

Module 1: Twitter Extraction

Module 2: Preprocessing

Module 3: Classification

Module 4: Report generation

4.1 TWITTER EXTRACTION

Client can team up as interface between the Client and the structure. New Client need to make a record by giving the username and secret key, the selected Client can straight forwardly login and can go into the structure twitter look for space. In look for space Client can give the data and Client get the tweets from the twitter. To expel the tweets, first the affiliation should be developed with twitter account using the twitter API called twitter4j. By then make the twitter architect application in twitter design site. From the made application we get the client key, puzzle key, Access token and token riddle key. Using these keys and tokens, it is Configured and connected with twitter. In this API it contains various parameters to think and read from the TwitterFactory by using request look and need to keep up the inquiry recorded records in Query Result. Using getTweets system we can get the tweets, from which we can evacuate the tweet username.

4.2 PREPROCESSING

The isolated tweets are the preprocessed by emptying stop words, short shape and emojis. Every single futile word in the tweets, for instance, stop words are been removed. Each and every short edge will be supplanted with full words so it is sensible for each one of the customers. Emoticons are known as smileys, there are shifts sorts of smileys. For each smileys there are some eager suppositions in it, which the customer use to pass on in generously less requesting way anyway it isn't fundamental all the customer will know the significance everything considered. Thus, every one of the emoticons is supplanted with their specific importance.

4.3 CLASSIFICATION

Support Vector Machines rely upon the possibility of decision planes that portray decision limits. A decision plane is one that detaches between courses of action of things having unmistakable class cooperation's. A schematic case: pharmaceuticals and diseases. After the preprocessing the tweets are orchestrated into catchphrase related tweets. The words are perceived in perspective of the watchwords to describe the tweets. This vocabulary examination procedure is used to find the favored class from the sweeping number of tweets.

4.4 REPORT GENERATION

The predictions based on the tweets contain information that refers to any disease from Twitter real-time feeds. This means the number of tweets that may contain the conversations related to disease symptoms and health outcomes are calculated. The reports are generated according to the number of tweets containing disease related contents in a country.

5. WORKFLOW

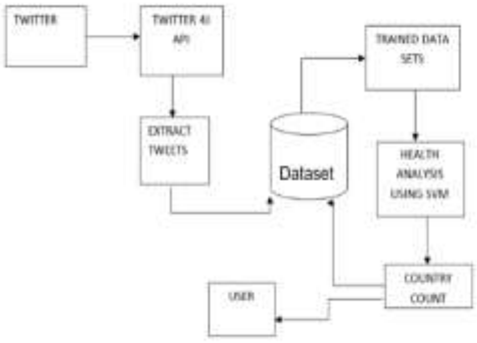

Figure 2: Workflow Model

6. PERFORMANCE LEVEL

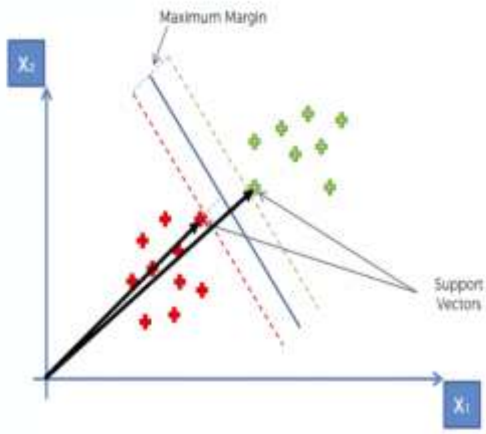

Figure 3: Efficiency level

7. CONCLUSION

On this paper, we've proposed a framework for ordering capsules in mild of extremity investigation of twitter facts. The twitter tweets are removed with twitter API utilising twitter4j. From the twitter created utility all the keys and token are produced, with these statistics we will associate the twitter with twitter API. At that factor extricated tweets are preprocessed through evacuating stop words, quick structures. The preprocessed tweets are characterised making use of Na� e Bayes grouping and extremity of the tweets is predicted for conclusive arrangement. This framework interpersonal business enterprise based social investigation parameters can build the forecast greater precision and speedy health examine.

8. FUTURE ENHANCEMENT

Our project can be enhanced with some features in future. Here we have not used the emoji's and symbols for sentimental analysis on the user tweets. In future we can consider those things along with the keywords so that it may increase the efficiency in a better level and also we can give a medication and remedy for those disease found in the country in future.

9. REFERENCES

1. L. Manikonda and M. D. Choudhury, "Modeling and understanding visual attributes of mental health disclosures in social media," in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017., 2017, pp. 170–181.

2. S. R. Chowdhury, M. Imran, M. R. Asghar, S. Amer-Yahia, and C. Castillo, "Tweet4act: Using incident-specific profiles for classifying crisis-related messages," in 10th Proceedings of the International Conference on Information Systems for Crisis Response and Management, Baden-Baden, Germany, May 12-15, 2013., 2013.

3. M. J. Paul and M. Dredze, "You Are What You Tweet: Analyzing Twitter for Public Health," in ICWSM'11, 2011.

4. Y. Wang, E. Agichtein, and M. Benzi, "TM-LDA: Efficient Online Modeling of Latent Topic Transitions in Social Media," in KDD'12, 2012, pp. 123–131.

5. S. R. Chowdhury, M. Imran, M. R. Asghar, S. Amer-Yahia, and C. Castillo, "Tweet4act: Using incident-specific profiles for classifying crisis-related messages," in 10th Proceedings of the International Conference on Information Systems for Crisis Response and Management, Baden-Baden, Germany, May 12-15, 2013., 2013.

6. C. X. Lin, Q. Mei, J. Han, Y. Jiang, and M. Danilevsky, "The Joint Inference of Topic Diffusion and Evolution in Social Communities," in ICDM'11, 2011, pp. 378–387.

7. X. Wang and A. McCallum, "Topics Over Time: A Non-Markov Continuous-time Model of Topical Trends," in KDD'06, 2006, pp. 424–433.

8. P. Barberá, "Birds of The Same Feather Tweet Together: Bayesian Ideal Point Estimation using Twitter Data," Political Analysis, vol. 23, no. 1, pp. 76–91, 2015.

9. L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter Sentiment Classification," in HLT'11, 2011, pp. 151–160.

10. S. Wang, M. J. Paul, and M. Dredze, "Exploring Health Topics in Chinese Social Media: An Analysis of Sina Weibo," AAAI Work World Wide Web Public Heal Intell, 2014.