Extraction and Analytics from Twitter Social Media with Pragmatic Evaluation of MySQL Database

Abhijit Bandyopadhyay
Teacher-in-Charge
Computer Application Department
Raniganj Institute of Computer and Information Sciences, Raniganj
West Bengal, India.

Abstract :

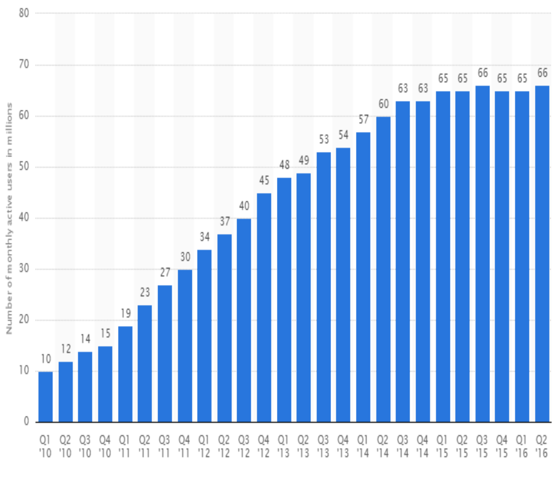
With the escalating use of web based platforms and technology loaded devices, a number of online communication environments are used. Social Media is one of the prominent as well as effective modes of communication for personal or group based broadcasting of emotions or sentiments. The process of communication and data transmission to like-minded or friends is traditionally known as social media in which a strong and performance aware web based environment is used as a media to deliver and transmit the emotions. The process of extracting and mining the data flowing on such channels is traditionally known as sentiment mining. This process is done using web based APIs which communicate with the live servers of social media. Now, the key point arises on the efficiency of database engine because the real time streaming data is very complex for a classical relational database management system. This research work focus on the implementation of Twitter Social Media Extraction and Sentiment Mining with the performance evaluation of MySQL Database System on multiple parameters. Twitter is one of the leading and prominent social media platforms that is used for the distribution, dissemination and broadcasting of views in multiple formats. A number of celebrities, political speakers, leaders and key personalities are using Twitter so that their views and sentiments can be transferred to the whole world. Even the media groups and news channels are using Twitter for the distribution of news in form of tweets to all the devices and handhelds. In this research manuscript, an effectual approach for the mining of social media tweets is presented to be used so that the understandable as well as prediction based popularity extraction can be implemented. The present work is based on the matching of positive and negative words from social media tweets which are proposed to be stored in a database engine so that the overall performance of database system with the real time data can be evaluated. The implementation aspects in this work shall use MySQL as back-end database in which the live tweets from Twitter Server based on a Java based platform. Any keyword can be searched on the Java based application that will communicate with Twitter Servers and the live streaming tweets shall be inserted in the MySQL Table. Using specific applications, plugins and programming interfaces, the information regarding particular keyword can be fetched and then the processing to be done from MySQL. The dynamic insertion of records in terms of real time streaming data from Twitter is projected to be done on different attributes of database including User Timeline, Tweets / Retweet, Number of Followers, Platform and Devices Used, Timestamp and Retweet Status, Friends' List, List of Retweets and Related Timezone, Retweet Platform and Followers, Individual Followers' Tree and many others. The key focus in this research work is to evaluate the efficiency factor associated with MySQL in handling the real time data from Twitter Servers based on the dynamic keywords to be processed in the Java based application.

Keywords: Twitter Database Interaction, Twitter Database Channels, Sentiment Extraction, Social Media Mining, Text Mining, Twitter Mining

Introduction

Sentiment Mining is one of the prominent domains of research in which the live prominence of different real time objects or persons can be evaluated. In this domain, enormous work is done based on the performance of text mining approaches, still there is huge scope of research in multiple segments including database performance, real time database congestions, concurrency issues and many others.

Twitter is widely used social media throughout the world and following is the statistical view from different perspectives and presents the statistics of Twitter used by assorted media and global locations. The following statistical data depicts the Twitter users of USA in year 2016.

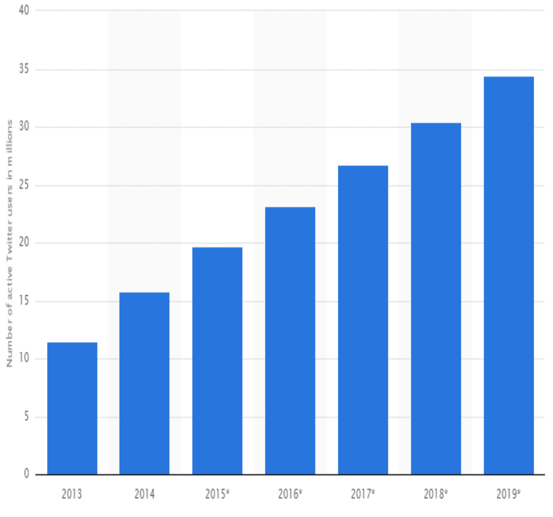


© Statista 2016

Figure 1. Statista Analytics on Twitter Data

The above depicted graph represents the escalating values of Twitter Users in USA and that is the key base of big data which is part of this research work in which the real time data analytics is done.

In another analytics by Statista, the world prominent statistical portal, following is the statistical Description of Twitter Users in India (2012-2019) in Predictive Aspects. The following statistical data and report depicts the forecasting and predictive analysis of Twitter users.



© Statista 2016

Figure 2. Statistical Description of Twitter Users in India (2012-2019)
Source : Statista

LITERATURE SURVEY

The extraction, analysis and learning of earlier work done on the related domain are very important aspect for any research work. To propose and defend the new and effective approach, there is need to present to the excerpts from assorted research papers with the deep learning of journals, conference proceedings and seminars. In this section, the extracts of earlier research work from assorted sources are mentioned so that the proposed work can be under investigation effectively.

K. Arun et al. underlines the sentiment analysis from Twitter Social Media on the specific theme of demonetization or currency transformation in India. The implementation on R language is found effectual as per the results from the work of authors. The approach of natural language processing and text mining is implemented in the work using Twitter APIs and writing of the fetched dataset into a file system. The generation of word clouds based on the thickness or density of the key words used is done with the comparison from multiple aspects including positive and negative keywords so that the scoring of tweets in terms of sentiments can be done. There are two phases in this work including accessing Twitter Data and Sentiment Analysis. The first phase comprises Authentication from Twitter using Security Tokens and Accessing the Data Sets from Twitter. The second phase is having data set extraction, text pre-processing, classification, scoring and plotting of results. The evaluation of 10000 tweets on demonetization, 2000 tweets from digital payments, 2358 tweets on operation clean money and 1983 tweets on income tax payments are extracted and evaluated with the performance aware results. The approach to use the open source platform of R is used so that the work can be done in effective manner without any issues of platforms.

S. Asur et al. presented the work on the predictive analysis using social media. The forecasting of revenues based on the tweets and sentiments extracts is done in the work with the demonstration of sentiments extraction from Twitter with the forecasting powers and functions on assorted perspectives. The key concepts of Subjectivity and Polarity are used for the accuracy of projected model. The results outperform the output obtained from Hollywood Stock Exchange on the similar aspects. The domain of sentiment scoring, recommendations and prediction is touched in this research manuscript by forms the foundations for development and deployment of a recommender system. In this research manuscript, the dataset from Twitter is fetched and then the recommendations about opinions or popularity of movies are implemented. The research manuscript presents the results in effectual aspects with the cavernous graphical representations having recommendations and predictions.

Bifet et al. presented in this research work with the key points of deep learning algorithms and machine intelligence approaches with the statistical analysis at the base level. Using the implementation of assorted statistical methods, there is specific predictive analysis in association with assorted data mining approaches. The work in this paper is presenting the effectual results with the analytical results with integration of assorted datasets in investigation. The work underlines the challenges associated with Twitter real time streaming data with the specific focus on the classification issues with the opinion mining. The work project the use of Kappa Statistics for the analysis of time based real time streaming data. The Kappa Statistics is used for the study and analytics on Twitter Data with the machine learning algorithms for evaluation of data streams. The challenges or key issues taken in this work includes community discovery and social information diffusion. The work on streaming data evaluation with the associated factors of unbalanced classes is done in this manuscript. The data stream mining approaches are multinomial Naive Bayes, Stochastic Gradient Descent and Hoedffding Tree which performs the implementation with higher optimization.

Jisha S. presents the work on Twitter emotions or opinion mining with the specific case of subject identification. The particular domain of subject identification is underlined with the analysis of real time tweets from Twitter servers. The proposed model in this work includes Crawling, Testing of Data, Data Preparation, Fisher Sentiment Analysis, Ranked Topics and the document probabilities. The simplest model of Bag of Words is projected in this paper. This model follows the approach of ignoring the ordering or words which discard the structure of sentence with the count and occurences of the words. The live fetched data is inserted into MySQL database engine with multiple attributes including, Text, Timestamp, User Information, Sentiment and Description.

Efthymios K. et al. evaluates the internal aspects of semantics as well as syntactics associated with the real time streaming Twitter data. The work focus on the analysis of linguistic feature points of tweets so that the detection of sentiments with each tweets and finally cumulative score can be done with higher degree of accuracy and performance. The hashtagged data Set (HASH) is used from Edinburgh Twitter Corpus with the data set of emoticons (EMOT) which are made available for research and development. The dataset of ISIEVE is used for the evaluation. From the results, it is found evident that n-gram+lex+twit based approach is performing better as compared to the traditional approaches of individual n-gram, n-gram-lex, n-gram-POS or all.

Varsha S. et al. worked on the sentiment extraction, scoring and predictive analytics is done which is now days very popular and prevalent in social as well as classical media platforms. In this research based approach, the predictive analytics show by which dimension the opinion from user profiles towards a particular event can be found and then overall scoring can be done. The work is having key focus on the Parts of Speech (POS) for the specific prior polarity features. The machine learning methods are covered in this work which can be used for greater efficiency. The methods mentioned in this paper includes Naive Bayes, Maximum Entropy and Support Vector Machines (SVM). The projected novel approach in this work is having multiple phases of implementation which includes retrieval of tweets, pre-processing of the live extracted data, parallel processing, sentiment scoring module and output sentiment evaluation.

Saif H. et al. is focused and diverted towards the market predictions and forecasting with the extraction of live dataset from social media. Using the live dataset fetching from Twitter and other social media profiles, the research work can be used to present the effective market survey and then product based prediction in terms of their related popularity, scoring and user sentiments. The projected approach in this work is having accuracy of 86% which is outperforming the earlier approaches worked on the similar domain. The sentiment classification results in this work are found quite effectual and effective with the higher degree of accuracy to 84% at highest level with the sentiment interpolation. In this research work, the authors present their work with effective and novel models. The initial approach of this research work is more diverted towards the extraction of live microblogs and tweets then training in association with the supervised model so that predictions can be done. This work is done with the approach similar to artificial neural networks. The second model of this research work present the predictive results in forms of classification of emotions, tweet sentiments and microblog opinions.

Saeideh S. et al. worked in this paper with the implementation in three different but integrated layers which performs the predictive analysis from user tweets fetched from Twitter social media platform. The first phase in this work fetches the real time streaming dataset from Twitter for training of the upcoming layers. The second layer in this work develops and implements the effective model with Naïve Bayes for classification of sentiments. The third layer or phase of this work is having focus on the testing and predictive presentation of results. The work is having SAS Architecture with multiple modules and phases including Training Corpora, Positive or Negative Sentiment, Statistical Model, Polarity Keywords and Final Testing. With the integration of SAS model for the classification of tweets, the results are outperforming earlier work. The approach is using Tree Tagger for the Tweets is also used with efficiency aware results.

Sanket P. et al. integrate assorted perspectives in Twitter mining including data preprocessing, sentiment analysis and NER. The proposed approach in this work includes removal of special characters, slang words, removal of stop words, Name Entity Resolution (NER) with the deep evaluation of the messages and emotions associated with the tweet. This research work is having key focus on the development and deployment of a new and effective classification model having analysis of popularity dataset simply known as word of mouth dataset. Using this approach, the words associated with positive and negative sentiments can be classified and further predicted. In this research work, the authors present the implementation aspects of social media platform with their related performance factors. The social media platforms are deeply analyzed and depicted the methods by which the extraction of user profile can be accomplished. Using this approach, there is clear dimension towards the social media profiles and their relationships which lays the foundation of base for further analytics or rule mining. The work develops an application in which the data file can be trained and then the keywords matching can be done. With the use of NER approach, the overall sentiment mining and opinion analytics is done. The steps of filtering, tokenization, removal of stop words and construction of n-grams are adopted in the construction of grams which is one of the key points in this research work.

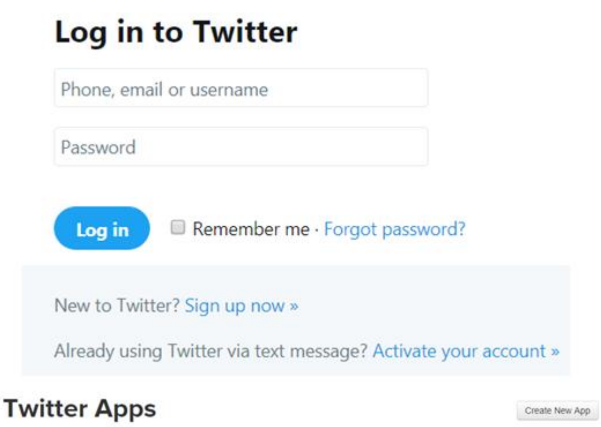
Anita B. et al. focus on the analysis of data management with the query handling with specific scenario of social networks and propose the use of NoSQL database for greater efficiency. The work focus on the use of different types of NoSQL databases as compared to traditional relational database management systems so that real time streaming data can be stored and fetched with minimum delay and higher optimal results can be achieved. The work propose the use of different paradigms and taxonomy associated with such databases including Wide Column Store (Column Families), Document Oriented database, Graph database and Key Value (Tuple Value) databases.

Proposed Work

- Fetching live data feeds from social media
- Data Cleaning or Refinement
- Identification of Feature Points in terms of Emotions based Words
- Database Connection for Live Storage and Fetching
- Extraction of the Mandatory Aspects
- Implementation of the Algorithm for Predictive Analytics
- Insertion of Records in MySQL Database
- Investigation of Positive and Negative Tweets
- Analysis of Popularity Score or simply Opinion Mining
- Detailed Analytic Report and Evaluation of Database Engine

Twitter Extraction Process

Following is the process of extracting the live tweets from Twitter using Twitter API and Developer Account



Log in to Twitter

Phone, email or username

Password

Log in □ Remember me · Forgot password?

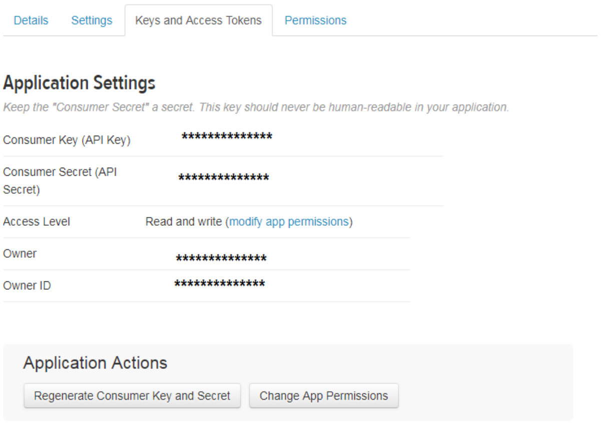
New to Twitter? Sign up now »

Already using Twitter via text message? Activate your account »

Twitter Apps

Create New App

Figure 3. Creating New App in Twitter



Details | Settings | Keys and Access Tokens | Permissions

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

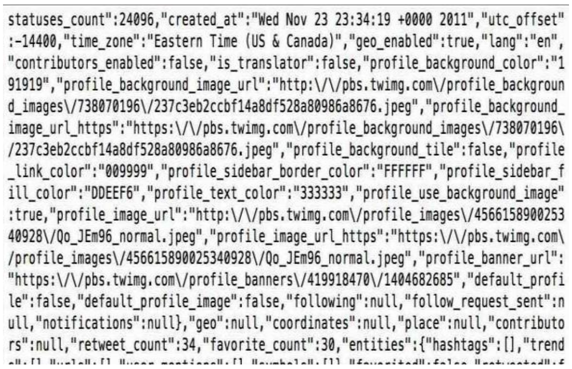
Consumer Key (API Key)	**************
Consumer Secret (API Secret)	**************
Access Level	Read and write (modify app permissions)
Owner	**************
Owner ID	**************

Application Actions

Regenerate Consumer Key and Secret | Change App Permissions

Figure 4. Generation of Authentication Tokens from Twitter

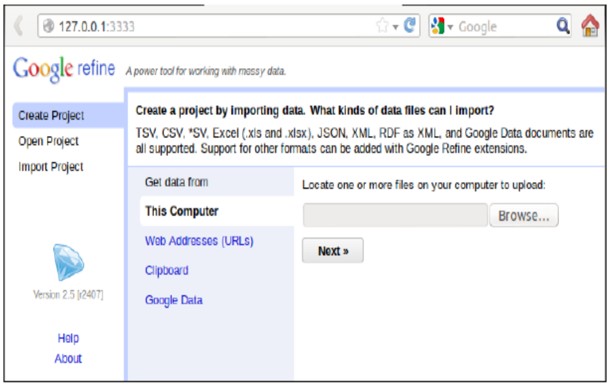
The abovementioned figure is the depiction of the authentication tokens and secret keys delivered by Twitter Platform. Without these tokens and keys, the live streaming data cannot be downloaded or checked. Using this approach, the platform of Twitter come to know about the particular person or organization willing to check the identities or tweets of other persons.



```
statuses_count":24096,"created_at":"Wed Nov 23 23:34:19 +0000 2011","utc_offset"
:-14400,"time_zone":"Eastern Time (US & Canada)","geo_enabled":true,"lang":"en",
"contributors_enabled":false,"is_translator":false,"profile_background_color":"1
91919","profile_background_image_url":"http:\/\/pbs.twimg.com\/profile_backgroun
d_images\/738070196\/237c3eb2ccbf14a8df528a80986a8676.jpeg","profile_background_
image_url_https":"https:\/\/pbs.twimg.com\/profile_background_images\/738070196\
/237c3eb2ccbf14a8df528a80986a8676.jpeg","profile_background_tile":false,"profile
_link_color":"009999","profile_sidebar_border_color":"FFFFFF","profile_sidebar_f
ill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image"
:true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/456615890253
40928\/Qo_JEm96_normal.jpeg","profile_image_url_https":"https:\/\/pbs.twimg.com\
/profile_images\/456615890025340928\/Qo_JEm96_normal.jpeg","profile_banner_url":
"https:\/\/pbs.twimg.com\/profile_banners\/419918470\/1404682685","default_profi
le":false,"default_profile_image":false,"following":null,"follow_request_sent":n
ull,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributo
rs":null,"retweet_count":34,"favorite_count":30,"entities":{"hashtags":[],"trend
```

Figure 5. Fetching Live Tweets from Twitter in JSON Format

The above depicted figure is the representation the execution of Code and fetching the results in JSON format (JavaScript Object Notation). The code provides the output in JSON format because JavaScript Object Notation is the generic or typical format of the live streaming big data.



Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with Google Refine extensions.

Locate one or more files on your computer to upload:

Figure 6. Parsing of JSON using Google Refine

Google Refine or Open Refine is the free and open source tool used for parsing or transformation of JSON to the understandable and formatted aspects so that the data mining and machine learning algorithm can be implemented. Earlier version was Google Refine which was further transformed to Open Refine under the URL openrefine.org so that the research community including scientists, academics and research scholars can use it.

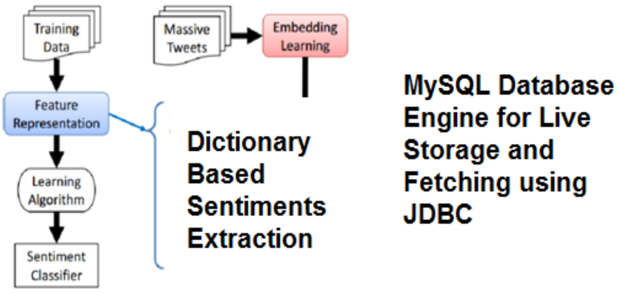


Figure 7. Projected Approach with MySQL Database Engine

The above mentioned figure represents the flow of sentiment classification. The process of sentiment classification solely depends on the bag of words in which positive and negative words are processed underweight and then finally checked so that the overall scoring of tweet regarding particular celebrity, person, entity or organization can be done. To extract the live data from Twitter Social Media, the APIs of Twitter4j are used which provides the programming interface to connect with Twitter Servers. In Eclipse IDE, the code of Java can be programmed for predictive analysis and evaluation of the tweets fetched from real time streaming channels. As social media mining is one of the key segment of research for extraction and prediction of popularity, the following code snippets are used to extract the real time streaming and evaluation of user sentiments.

Database Structure with Different Fields

On life fetching the data from Twitter Servers, the following database fields can be accessed and stored to the database table so that the overall performance in the storing as well as retrieval can be done.

User Timeline
Tweets / Retweet
Number of Followers
Platform and Devices Used
Timestamp and Retweet Status
Friends' List
List of Retweets and Related Timezone
Retweet Platform and Followers
Individual Followers' Tree

The proposed model is an effective and performance aware approach for opinion mining and the analysis of user timeline from assorted social media so that a common platform or application do not repeatedly require the sign on. Using this approach, the user can be identified on the heterogeneous media platforms and identity can be mined. Once the identity of user is mined, the further creation and activation of new account will not be required. Using this approach, the performance, complexity and time can be optimized a lot with huge optimization factors. The work begins with the experimentation done on the fetched tweets according to the categories. In addition thorough analysis of different tweets is done using deep mining and association of tweets and tokens. Further the proposed technique is being compared with the existing techniques.

The proposed model is an effective and performance aware approach for opinion mining and the analysis of user timeline from assorted social media so that a common platform or application do not repeatedly require the sign on.

Using this approach, the user can be identified on the heterogeneous media platforms and identity can be mined. Once the identity of user is mined, the further creation and activation of new account will not be required. Using this approach, the performance, complexity and time can be optimized a lot with huge optimization factors.

The different stages for this work are followings:

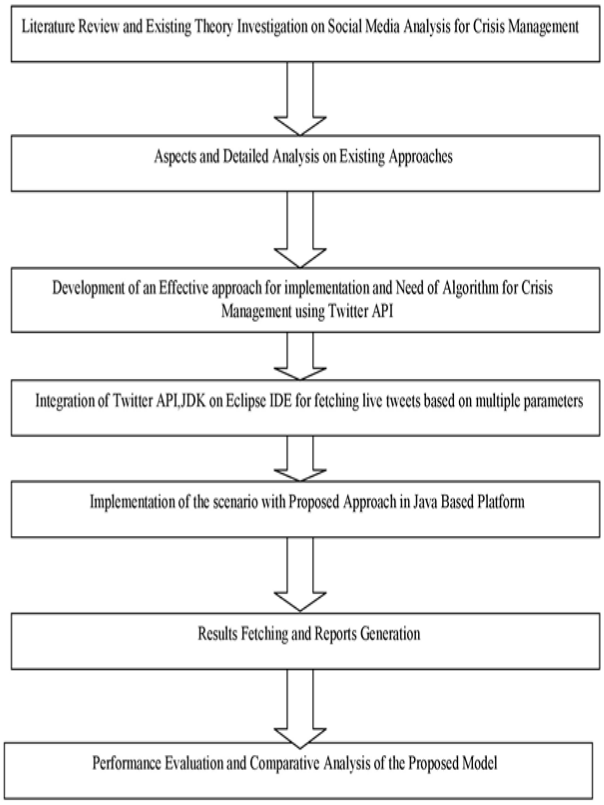

Literature Review and Existing Theory Investigation on Social Media Analysis for Crisis Management
Aspects and Detailed Analysis on Existing Approaches
Development of an Effective approach for implementation and Need of Algorithm for Crisis Management using Twitter API
Integration of Twitter API,JDK on Eclipse IDE for fetching live tweets based on multiple parameters
Implementation of the scenario with Proposed Approach in Java Based Platform
Results Fetching and Reports Generation
Performance Evaluation and Comparative Analysis of the Proposed Model

Figure 8. Steps of the Proposed Research Work

Tweets	Execution Time (ms)
10	0.899
20	0.977
50	1.8
100	1.988

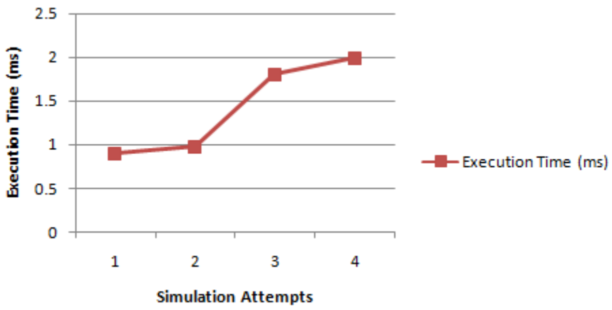


Figure 9. Execution Results and Evaluation of Time

CONCLUSION

In this research work, the live extraction of timeline from social media platforms is implemented with the multidimensional performance evaluation of the database system with respect to real time streaming data. The work presents an approach by which the effectiveness of MySQL Database Engine in handling the real time data so that the suitability of database systems can be proposed. The recommendation associated with this work is towards the predictive mining on assorted events, celebrities or popularity factors in real time domain. The predictions associated with business including stock market can be done effectually with this implementation. There are number of optimization approaches using which the efficiency, accuracy and performance factors can be improved. The integration of soft computing approaches are prevalent in the research community which provides fuzzy based execution and global optimization from existing results. For future scope of the work, the soft computing techniques can be used in hybrid approach with the analytics of MySQL to have better and efficient results in terms of higher degree of optimization. As MySQL database is the key technology or reservoir for handling the live streaming data, still there exist huge scope of research. The concurrency issues with the functions and triggers associated with MySQL can be analyzed effectively in future work so that optimality can be achieved with minimum error rate.

REFERENCES

[1] P. Varshawangikar and P. K. Jayamalini, “Data Preprocessing , Sentiment Analysis & NER On Twitter Data .,” pp. 73–79.
[2] K. Arun, A. Srinagesh, and M. Ramesh, “Twitter Sentiment Analysis on Demonetization tweets in India Using R language,” Int. J. Comput. Eng. Res. Trends, vol. 4, no. 6, pp. 252–258, 2017.
[3] S. Asur and B. a Huberman, “Predicting the Future with Social Media,” Computing, vol. 1, pp. 492–499, 2010.
[4] S. Shahheidari, H. Dong, and M. N. R. Bin Daud, “Twitter sentiment mining: A multi domain analysis,” Proc. - 2013 7th Int. Conf. Complex, Intelligent, Softw. Intensive Syst. CISIS 2013, pp. 144–149, 2013.
[5] E. Kouloumpis, T. Wilson, and J. Moore, “Twitter sentiment analysis: The good the bad and the omg!,” Proc. Fifth Int. AAAI Conf. Weblogs Soc. Media (ICWSM 11), pp. 538–541, 2011.
[6] T. M. S. Akshi Kumar, “Sentiment Analysis on Twitter,” Int. J. Innov. Res. Adv. Eng., vol. 2, no. 1, pp. 178–184, 2015.
[7] C. Engineering and J. S. Manjaly, “Twitter based Sentiment Analysis for Subject Identification,” vol. 2, no. 5, pp. 2026–2029, 2013.
[8] A. Bifet and E. Frank, “Sentiment Knowledge Discovery in Twitter Streaming Data.”
[9] A. B. Mathew and S. D. Madhu Kumar, “Analysis of data management and query handling in social networks using NoSQL databases,” 2015 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2015, vol. c, pp. 800–806, 2015.
[10] H. Saif, Y. He, and H. Alani, “Alleviating data sparsity for twitter sentiment analysis,” CEUR Workshop Proc., vol. 838, pp. 2–9, 2012.