ARTICLE

AN ASSESSMENT MODEL TO EVALUATING THE QUALITY OF SOCIAL MEDIA DATA USING QUALITY ATTRIBUTES FOR IMPROVING BUSINESS DECISION MAKING

Supriya Haribhau Pawar, Devendrasingh Thakore

Bharati Vidyapeeth Deemed University College of Engineering, Pune, Maharashtra, INDIA

ISSN 0976-3104

ABSTRACT

Background: *Big data describe volume amount of structured and unstructured data. Big data sources like social media, Facebook collected large number of unstructured data. When unstructured data collected from different sources, maintain the quality of data is also important. Big data sources provide insights to businesses for improve business decision making. The proposed system improves the quality of business data which are collected for business decision making. Data quality Assessment is a way for practitioners to understand the scope of how poor data quality effects on business process and develop a business case for data quality management.* ***Methods:*** *This paper contributes to providing a solution by introducing new assessment model to evaluate and manage the quality of social media data. Sentiment analysis is used for monitoring real-time data. Generate the new rules and attributes to assess the quality of data. Apply quality attributes on input data and assess only those data which are fit into the quality attribute dimensions. Evaluate data quality using large data set.* ***Result:*** The system provides the visualized data and generates a report based on sentiment analysis. ***Conclusion:*** *The proposed system improve solution by provide real time and validate data for the user.*

KEY WORDS
Data quality, social media data, Sentiment Analysis, Quality attributes, evaluating quality attributes.

Received: 23 June 2017
Accepted: 29 Aug 2017
Published: 20 Oct 2017

***Corresponding Author**
Email:
pawar45supriya@gmail.com
Tel.: +91+8451930918

INTRODUCTION

Big data is a term that describes large volume of data. Data are both unstructured and structured format. Now day's large number of data are generated in business and this data important in business point of view. Analyzing large data set in big data architecture is most important. Analytics process in big data has the capability to disclose patterns, movement and associations. Especially process of big data affecting to businesses to achieve the goal of decision making. Extracting information from large data set is very important. For extracting large data set carefully do the planning and provide relevant input to decision layers. When adopting solution for big data, businesses are needed to use new technology and framework for large data set.

Social media data is one of big data's most influential origins. The big data dimensions: Volume, variety, velocity, and veracity, produce some challenges not only to data analytics but also to the data system that manage the data. In big data system, input data is eventual source of knowledge. In big data, lifecycle data travel in four phases as shown in [Fig. 1]: data generation, data acquisition, data storage and data analytics. In data generation phase data are created, data are created from larger number of sources, for example, social media sites, Facebook. Data acquisition phase consist of collection of data, transmission of data and pre-processing of data. Generated data are in structured, semi-structured and unstructured format. After collecting data cleansing, data deduplication, filtering is done in preprocessing phase. When data is collected form data sources maintain the data quality is also important. Improve the quality of data by applying quality attributes on input data. Quality attributes are accessed, controlled and improve the data that impact the result of the analysis phase. After data preprocessing data are stored for analysis the results.

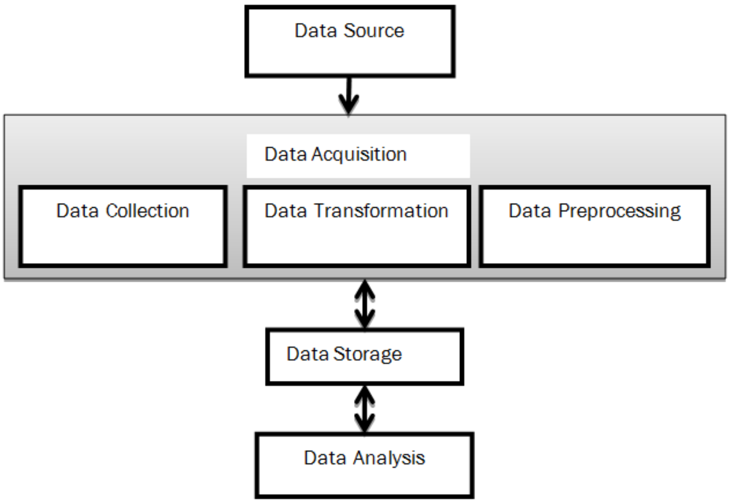

Fig. 1: Lifecycle of big data.

………

Social media platform offers marketers significant amount of data which can be used to make a marketing decision in the future [1]. Goals of social media data are an improvement of customer service, instant feedback on products and services through sentiment analysis. Use different metrics to reach different goals. When Social media data collected from various sources (Facebook, twitter), common challenges are arises in data like missing or incomplete data, unavailable of data streams, old data. The user wants to ensure the reliability of the data while collecting. When data is proceeding for analyzing some data, the user intends to make sure that the relevancy and quality of data are appropriate the particular solution. Reliable and valuable data enhance decision making of business. The evaluation of data quality happens in data processing phase in big data architecture, data extraction, data processing, and decision making. Quality evaluation of big data considers while data goes through the pipeline of big data system [2].

When unstructured data collected from different sources, maintain the quality of data is also important. Unstructured data (Input data) are important for marketer for taking a right decision and gain valuable customer insights, reduce marketing expenses and improve sales. The proposed system ensures the quality and trustworthiness of social media data for business and marketers decision-making [3]. Data quality management is one where evaluate data quality and improve business decision making. Data quality management is continuous analysis, observation and improvement overall quality of the organization. The purpose of this paper is to how to evaluate the quality of data and provide trustworthiness data to business for decision making. Introduce solution for data evaluation, in which data customer can select relevant quality attributes and metrics and evaluate quality attributes with evaluation metrics.

LITERATURE SURVEY

Big data and big data quality

Big data

Big data is a term that describes large volume of data. Data are both unstructured and structured format. Unstructured data like document files, social media data, and website content. Structured data like SQL database stores. Big data is large data set and use this data to analyze computationally association, trends and human interaction and behavior.

Data quality

Data quality is a process of assessment of data. It defines set of values to gather qualitative and quantitative data. Data is consider high quality if data use in decision making, planning and operation. Quality of data is measure using attributes of data quality.

Quality attributes

Data quality dimension is term used to in business to evaluated data quality. Associations select the information quality measurements and related measurement edges in light of their business setting, prerequisites, and levels of hazard and so on.

Data quality attributes might be used to:

1. Recognize which information things should be evaluated for information quality, normally this will be information things considered as basic to business operations and related administration detailing.

2. Survey which information quality measurements to utilize and their related weighting.

3. For every information quality measurement, characterize values or ranges to great and quality information.

4. Apply the appraisal criteria to the information things.

5. Audit the outcomes and decide whether information quality is worthy or not.

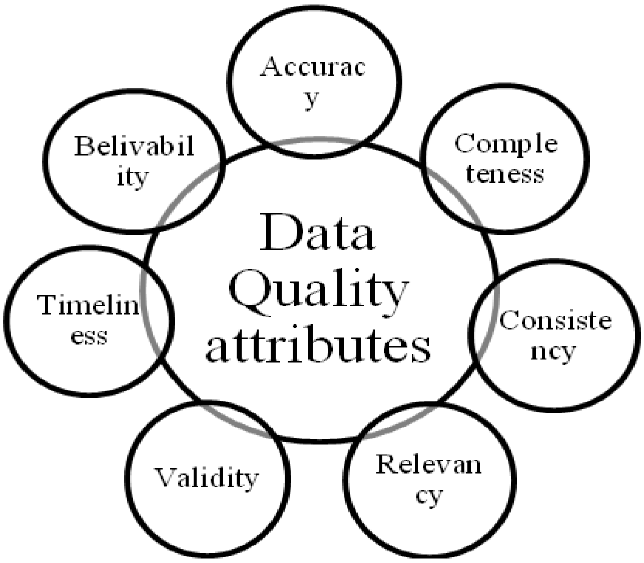

Fig. 2: Data quality attributes.

……….

Quality attributes definition [Fig 2]

1. *Accuracy*: Ensure that the input data is error free
2. *Completeness*: Check the extracted data is not missing
3. *Consistency*: Implies that not two or more values conflict with each other
4. *Relevancy:* The extracted information is helpful for the task. Non relevant data should not be considered.
5. *Validity*: Input data is valid in its purposed used.
6. *Timeliness:* The extracted input data is not old data. The timestamp is necessary when retrieving the data.
7. *Believability:* the extracted data is valid and credible

Quality metric

Quality metrics are components of an effective quality management plan and measure properties of quality attributes. To assess and analyze the quality in any system, first need to characterize any quality attributes which are relevant to that system. In proposed system quality attributes are consider:

- Relevancy: The extracted information is helpful for the task. Non relevant data should not be considered.
- Timeliness: The extracted input data is not old data. The timestamp is necessary when retrieving the data.
- Believability: the extracted data is valid and credible.
- Accuracy: Ensure that the input data is error free

RELATED WORK

Big data is relevant to many components like government, healthcare, business management, social media, education, life science. Using big data these components improve their decision making, transparency, quality by providing continuous monitoring. The challenges are arises when the volume of structured and unstructured data coming from different sources. The reason for generating a large amount of data, big data-based application introduced new challenges and issues for quality assurance engineers [4]. These challenges are not only limited to data analysis but also to a big data system that manages all the information [1].
Recently social media data such as Twitter, Facebook increase business by providing insights into customer opinions, thoughts, and preferences. Design the platform that supports to monitoring and analyzing customer feedback in social media network and identifies issues which are faced by customers. Internet users communicate and express their thoughts with thousands of other people. People use a social media platform to share their ideas and experiences with different customer products and services [2].

When data are coming from different sources maintaining or evaluating the quality of data is also important. Quality metrics are components of an effective quality management plan and measure properties of quality attributes. To assess and analyze the quality in any system, first need to characterize any quality attributes which are relevant to that system. Quality attributes such as Accuracy, performance, consistency, timeliness, completeness, relevancy [3] [4] are used to evaluate quality in social media data. Research aim [4] is providing

THE IIOAB JOURNAL

COMPUTER SCIENCE

trustworthiness metrics for information provenance and quality evaluation.Quality metrics measure the performance of product and processes. Each metrics has following properties [2].

- Description: Metric description.
- Purpose: The purpose of metric.
- Target: Where metric are used.
- Formula: How the value of metric is achieved.
- Range value: Value of range for the metric evaluation.
- Acceptable Value: Minimum value for accepted quality attributes.
- Rules: The set of measurement value range and a set of constraints which define a set of target measurement.

The above observations and literature studies [1-6] indicate that quality evaluation is limited to only a few qualities attributes, the purpose is to increase the quality assessment to introduce more quality attributes. Data quality management is one where evaluate data quality and improve business decision making. Data quality management is continuous analysis, observation and improvement overall quality of the organization. Unstructured sources like social media need data quality management for improving their data quality. Data quality management is one where evaluate data quality and improve business decision making. Data quality management is continuous analysis, observation and improvement overall quality of the organization. Ensuring data quality involves following steps [Fig 3]:

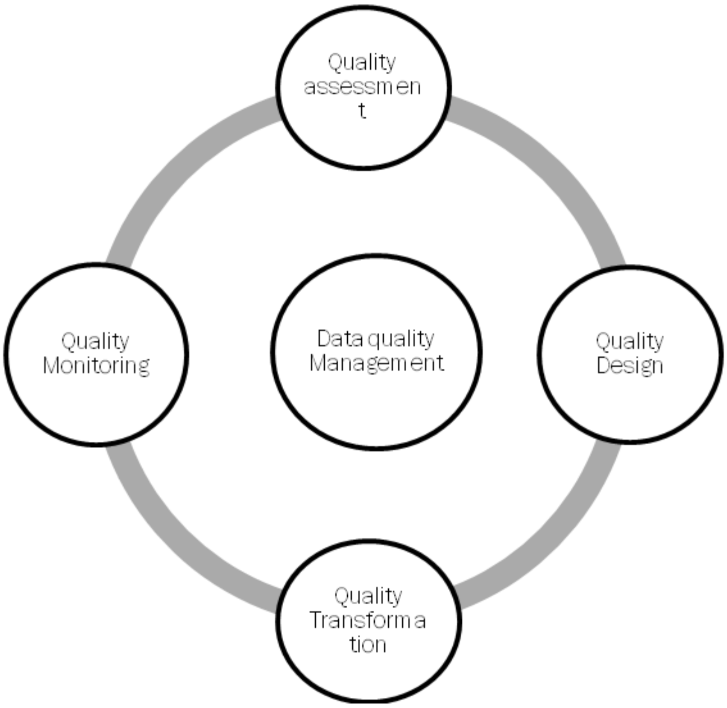

Fig.3: Data quality management.

……………………………………………………………………………………………………..

Quality assessment- In quality evaluation phase, decide data source type to extract the data. It is a way for practitioners to understand the scope of how poor quality data affects on business process and develop a business case for data quality management.

Quality design

In the quality design phase, design and analysis the quality process and concentrate on the data elements that are consider based on the selected business user needs.

Quality transformation

In quality transformation phase, define business related quality rules and perform measurement using metrics

Quality monitoring

In quality monitoring phase, review expectations and refine rules and monitor data quality versus target.

IIOAB JOURNAL ISSN: 0976-3104

MATERIAL AND METHODS

Architecture of proposed system is given in Fig-4.

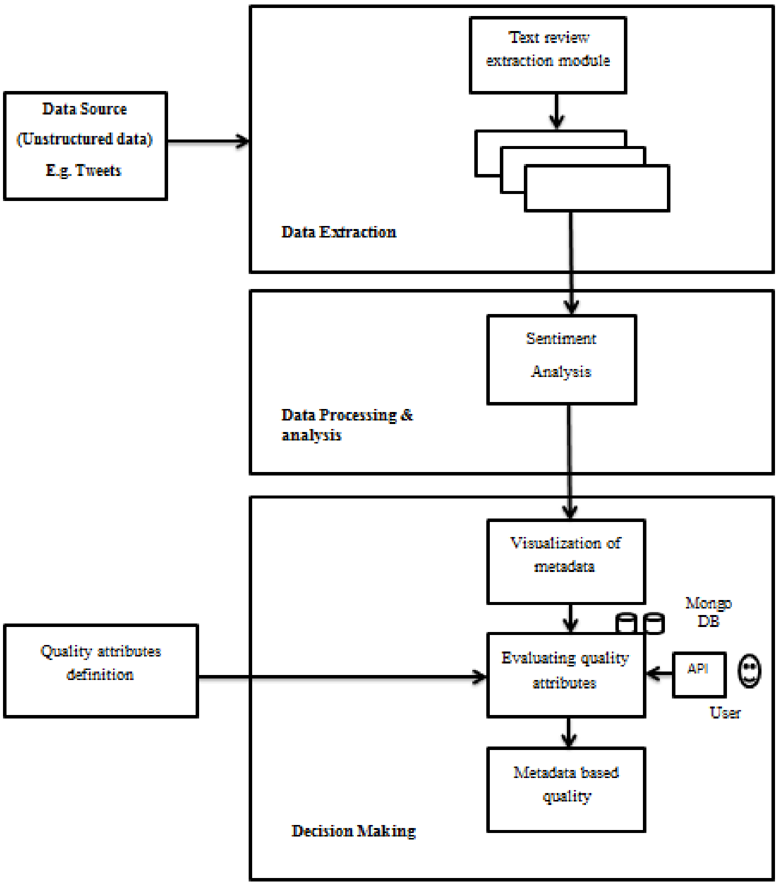
Data Source
(Unstructured data)
E.g. Tweets
Text review extraction module
Data Extraction
Sentiment Analysis
Data Processing & analysis
Visualization of metadata
Mongo DB
Quality attributes definition
Evaluating quality attributes
API
User
Metadata based quality
Decision Making

Fig.4: Architecture of proposed system.

Data Source

The main purpose of the system is to collect social media data (e.g. tweets) to achieve customer insight that can be used in business decision making. The decision-making policies have a great importance in quality evaluation. Social media source like tweeter where peoples are connecting to share their opinions and find out what happening in the world right now (Input data). Input data (unstructured data) are useful to business for improve data quality and decision making.

Data Extractor

Data extractor extracts tweets or document. Input data comes in preprocessing phase where data gets filtered. The links and punctuations are removed in that data and data become relevant to the system. Extracted data are in .json format. The design has been presentedfrom utilizing the developed tool for quality management of Twitter or .json based data sets. First, parse .json file by following attributes: Battery, Charger, display, phone, price, processor, ram, rare camera, screen, SD card, and weight.

Quality attributes

Apply quality attributes on input data and assess only those data which are fit into the quality attribute dimensions. Sentiment analysis is applied on assess data. Sentiment analysis is the process of recognition whether extracted data is positive, negative or neutral. Input data is scanned for positive or negative words like sad, happy, great, and terrible. Algorithms are used to score the document to decide whether they indicate positive or negative sentiment. Evaluate The Sentiment Analysis by Review Text Polarity and Polarity Confidence. Store the Polarity Confidence in Unstructured Database MongoDB which store data as documents. Then applying Decision Making Quality Attributes is:Accuracy, Timeliness, Confidences, relevance, popularity Tweets are created based on metadata, and related quality attributes. Metadata are a help to users to validate the quality and value of data for business usage. Manage quality metadata and attributes; rules are needed to define in quality, i.e. which quality attributes can be used and where. The system applies below Quality attributes [5].

- *Accuracy*-Ensure that the input data is error free
- Timeliness- The extracted input data is not old data. The timestamp is necessary when retrieving the data.
- *Confidence* – How quality based data is important for business decision making is decide by confidence attribute.
- *Relevance*- The extracted information is helpful for the task. Non-relevant data should not be considered.
- *Popularity*- Source of the system provides correct information and this information having numbers of followers.

Visualization of metadata

The relevant data is visualized to the end user, and decision-making policy defines the valuable data for decision making by selecting only those data sets with the correct quality attribute value.

Sentiment analysis algorithm

Input is text string. Input is always in JSON format. JSON syntax is a subset of JavaScript syntax. The JavaScript function JSON.parse(text) can be used to convert a JSON text into a JavaScript object. It involves the breaking down text into component parts with explanation of function and syntactic relationship of each component part. . First load positive words and negative words from text. If in text find positive word then count of positive word is incremented by 1 or if find negative word then count of negative word is incremented by 1. After finding positive and negative words calculate positive ratio and negative ratio. That is positive ratio is count by positive count divide by count of all words and negative Ratio is count by negative count divided by count of all words. Positive ratio set text is applicable and negative ratio set text is not applicable.If polarity is Applicable Then find positive Count in text and Apply range in between 0 -5.

Algorithm 1: Sentiment analysis

```
Input- text string
Output- Confidence
    1.  positiveWords = load positive words
negativeWords = load negative words
    2.          for each tweet:
parse the tweet
    3.          tweetWords = all the words in the tweet text
         positiveCount = 0
         negativeCount = 0
    4.          if candidate is in the text:
       if a positive word is in the text:
        positiveCount = positiveCount + 1
      if a negative word is in the text:
        negativeCount = negativeCount + 1
    5.          positiveRatio = positiveCount / count of all words
        Polarity = Applicable
negativeRatio = negativeCount / count of all words
        Polarity = Not applicable
    6.  Confidence: If polarity is Applicable Then find positiveCount in text and Apply range in between 0 -
        5
        ElsePolarity is not applicable
```

A. *System Algorithm*

Algorithm 2: System Algorithm

```
Input= data set
Output= complete data set

1. Set of input R= {r1, r2, r3,………., rn}
   Where,
    rn = total number of reviews
```

2. Parse document Pd= Pd(r1,r2,r3,………rn)= json(r1,r2,r3,………rn)
   Where,
   Pd = total numbers of reviews parsed

3. Text extraction
   Fp= {f1, f2, f3,………fn}
   Where,
   Fp= product features
   Therefore,
   Text extraction= $\sum_{f=1}^{n}(Pd\{r1,r2,r3 \dots \dots\dots rn\})$
   Tr= $\frac{|\{Fp\}\cap\{Pd\}}{|\{Pd\}|}$

   Where,
   Tr= Total number of reviews extracted from features of product
   Fp= Product Features
   Pd= total numbers of reviews parsed

4. Let, W= Wordcount
   WTq= bw (Tq1, Tq2, Tq3……..Tqn) and gw (Tq1, Tq2,Tq3……..Tqn)
   Where,
   bw= badword
   gw= goodword

   WTq are the words count of bw and gw for finding support.

5. Quality attribute
   Set of quality attribute Q= {q1, q2, q3, q4}
   Tq= $\sum_{q=1}^{n}(Tr\{tr1,tr2,tr3 \dots\dots\dots trn\})$
   Where,Tq= Total number of reviews extracted from quality attributes

6. Supp (Tq)= $\frac{t\epsilon gw;Tq\subseteq t}{|gw|}$
   t= data set
   gw= good words
   Tq= support

$\mathrm{Confidence}(Tq \rightarrow gw)$ = $\frac{supp(Tq\cup gw)}{supp(gw)}$

7. Accuracy (Tr) = $\frac{\sum_{Tr=0}^{Tr=n} Tr1}{\sum_{Fp=0}^{Fp=n} Fp1}$

   Label= Complete data or Incomplete data

B. Mathematical model

Mathematical model

---

1. Tr =$\frac{|\{Fp\}\cap\{Pd\}}{|\{Pd\}|}$

   Where,

Tr= Total number of reviews extracted from features of product
Fp= Product Features
Pd= total numbers of reviews parsed

2. Supp (Tq)= $\frac{t\epsilon gw;Tq\subseteq t}{|gw|}$
   t= data set
   gw= good words
   Tq= support

3. $\mathrm{Confidence}(Tq \rightarrow gw)$ = $\frac{supp(Tq\cup gw)}{supp(gw)}$

4. Accuracy (Tr) = $\frac{\sum_{Tr=0}^{Tr=n} Tr1}{\sum_{Fp=0}^{Fp=n} Fp1}$

RESULTS

The proposed system provides real time and quality data to business for decision making. Identify data items or input data need to be assessed for business data quality. Existing system taking decisionfor evaluate quality of data based on rating. If reviewer give the four rating to particular product and write comment product is not good that time businesses consider only rating not comment. From this analysis system businesses taking wrong decision. In proposed system taking decision based on comments. Apply quality attributes on input data and evaluate only those data which are fit into the quality attribute dimensions. For each quality, attribute defines the range to reprinting good or bad quality data. Using relevancy attribute businesses check for which product user said product is good or product is bad. For analysis, the result consider five reviews [Table 1]. First, apply timeliness attribute, the extracted input data is not old data. In proposed system consider only after 2014 year reviews. Then apply relevancy attribute, forextracted information is helpful for the task and in this situation,non-relevant data should not be considered. From relevancy find accuracy of data. Accuracy defineinput data is error free. Then find popularity of data. Form popularity attribute find which data is useful and which data is not useful.

Table 1: Results of five reviews

No. of Reviews	Confidence	Analysis
1	0.65497869	Incomplete data
2	0.90458458	Complete data
3	0.59680178	Incomplete data
4	0.98198456	Complete data
5	0.98235877	Complete data

Sentiment analysis is applying on usedfull data and finds confidence of the data. Sentiment analysis determines positive and negative sentiment from text. Sentiment analysis API (or document) provides very accurate analysis of the emotion of the text from sources. The analysis of text presented in range (e.g. range between 0 to 5). The result of scores closer to 5 considered to be positive sentiment and scores closer to 0 will be of negative sentiment. Based on above principle figure show the result is, 63% data is incomplete and 96% data are complete, which is shown in [Fig 5].

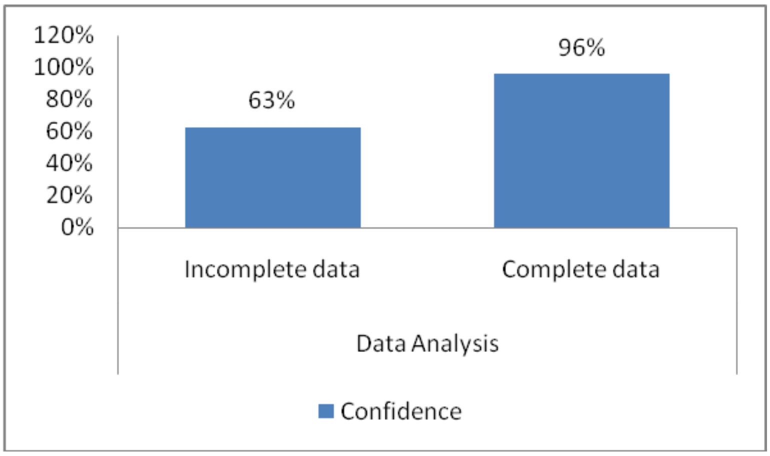

Fig. 5: Result of data analysis.

CONCLUSION

Most of existing research addresses different solutions to evaluate data quality in big data such as data cleansing using quality rules based on different functional dependency, data provenance. The quality of data is assessed in each of data processing phase. The proposed system provide preprocessing quality framework for big data quality. Generate quality rules which are applied on preprocessing activities to data analysis. Apply quality attributes on input data and evaluate data quality. The system analysis various data across the social web and helps to convert data into actionable insights. Sentiment analysis is used to review the social media data. The system provides the visualized data and generates a report based on sentiment analysis. Generation of data quality process on sample data set provides very faster result of data quality evaluation and instance updates are done when quality rules are inserted or deleted.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

To prepare proposed methodology paper on “An assessment model to evaluating the quality of social media data using quality attributes for improving business decision making" has been prepared by Supriya Pawar and Prof. Dr. Devendrasingh Thakore. I would like to thank my faculty as well as my whole department, parents, friends for their support. I have obtained a lot of knowledge during the preparation of this document.

FINANCIAL DISCLOSURE

None

REFERENCES

[1] Immonen A, Pääkkönen P, Ovaska E.[2015] Evaluating the Quality of Social Media Data in Big Data Architecture. IEEE Access. (2)3: 2169- 3536.

[2] Fabijan A, Olsson HH, Bosch J.[2015] Customer Feedback and Data Collection Techniques in Software R & D : A Literature Review. ICSOB 2015. 1: 139–153

[3] Immonen A, Palviainen M, Ovaska E.[2014] Requirements of an Open Data Based Business Ecosystem. IEEE Access. VOLUME 2, 2014. 2: 88-103

[4] Bhatia S, Li J, Peng W, Sun T. [2013] Monitoring and Analyzing Customer Feedback Through Social Media Platforms for Identifying and Remedying Customer Problems. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 3: 1147–1154.

[5] Bertino E. [2013]Big data - Opportunities and challenges: Panel position paper, Proc Int Comput Softw Appl Conf, (3) 6: pp. 479–480

[6] Nurse JRC, Rahman SS, Creese S, Goldsmith M, Lamberts K.[2011] Information Quality and Trustworthiness : A Topical State-of-the-Art Review. 2011 IEEE. (5)3: 492–500.

[7] Cong G, et al. [2007] Improving data quality: consistency and accuracy. Proceedings of the 33rd international conference on Very large data bases, 7: pp. 315–326.

[8] Tao C, Gao J. [2013] Quality Assurance for Big Data Application – Issues, Challenges, and Needs. SEKE 2016. (2)3:1-7.

[9] Bobrowski M.[1999] Measuring data quality, Univ. Buenos. 1428:99–102

[10] Poe S, Vrbsky SV. [2015] Comparing nosql mongodb to an sql db Comparing NoSQL MongoDB to an SQL D. Conference University of Alabama. 5:569-575.

[11] Kumar L, Rajawat S, Joshi K. [2015] Comparative analysis of NoSQL (MongoDB) with MySQL Database. International Journal of Modern Trends in Engineering and Research (IJMTER). (5)2: 120–128.

[12] Pawar SH. [2016] A Study on Big Data Security and Data Storage Infrastructure. International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE). (7)6:539–542

[13] Patil PT,[2016] A Study on Evolution of Storage Infrastructure. International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE). (7)6:501–506.

[14] Thakore D, Upadhyay AR.[2013] A Framework to Analyze Object-Oriented Software and Quality Assurance, International Journal of Innovative Technology and Exploring Engineering (IJITEE). 5:254–258.

[15] Devendrasingh Thakore, Akhilesh R Upadhyay. [2013] Analysing and Improving Quality Assurance in Document Search- Engine incorporating a Document-Ranking Algorithm for Text-Mining, International Journal of Latest Technology in Engineering, Management & Applied Science (IJLTEMAS). (5)2:17-23

[16] Thakore DM, Kadam SU.[2012] Increasing Scalability of Data Mining Algorithms for High Dimensional Data, International Journal of Advances in Computing and Information Research (IJACIR). (4)9:85-89

[17] Kabugade, Rohan R, Dhotre SS, Patil SH. [2014] A Study of Modified O (1) Algorithm for Real Time Task in Operating System. Sinhgad Institute of Management and Computer Application NCI2TM.(5)49: 223-230

[18] Karande Poonam, Dhotre SP, Suhas Patil. [2014] Task management for heterogeneous multi-core scheduling. Int J Comput Sci Inf Technol 5(1): 636-639.