Mining the Twitter Data Stream: A Review

MUHAMMAD USMAN*, MUHAMMAD AKRAM SHAIKH

Pakistan Scientific and Technological Information Center, Islamabad, Pakistan

**Corresponding Author's e-mail: usmiusman@gmail.com*

Abstract

Now a days, due to recent growth of the Internet and the World Wide Web large volume of data is available everywhere. Various approaches have been proposed in the literature to analyze such large volume data involving large number of dimensions. Real time data streams of large volume can be exploited for knowledge discovery. In the current study, a literature review of the different approaches adapted by various researchers to extract knowledge from Twitter data streams particularly by creating a multi-dimensional architecture comprising of Meta tags with tweets is conducted. These approaches have been reviewed in terms of data retrieval, storage, database design, analysis, visualization abilities and the technologies being used for these components. The study also discusses the technological gaps and proposes some recommendations for the future research.

Keywords: Twitter, Data Streams, Data Mining, Data Warehouse, Multidimensional Mining

INTRODUCTION

Over the last two decades, the data is being generated in large volume due to the use of automated systems, enhancements in storage capacity of devices and availability of internet connectivity [1]. Large volume of data contains important information related to specific domains and can be exploited further for knowledge extraction and high-level decision making. Many knowledge extraction techniques have been proposed by different researchers which help the decision makers to study the insights of data [2]. Data warehouses are used to store large volume of data in aggregate form to enable the multi-dimensional analysis. Moreover, data mining techniques have been used in the past to extract and predict trends by researchers [3]-[5].

Data streams are specific type of big data, in which real-time data is generated in large volume. Streaming data is available in all social networks like Twitter, Facebook and Instagram etc. at real time and in big volume. For example, according to a study [6], only Twitter generates 500 Million Tweets per day. Analysis of these data streams is gaining popularity these days.

In previous studies, researchers have exploited the Twitter Streaming API and stored the tweets in local file systems or databases for analysis purposes. Some studies have focused retrieving real-time data for real-time analysis, while others have saved data over a specific period of time for different analysis perspectives. Apart from this, the tweet text is targeted for textual and opinion mining along with the analysis which is done on Meta tags associated with each tweet. Data storage has been done in files systems, open-source databases, licensed database as well as aggregated storage repositories like data warehouses. The targeted objective in the research mainly include topic identification, location based intensity of Twitter usage, improvements in storage time of data streams, application of OLAP over multidimensional stream data, behavioral analysis, crisis detection and other events detections. Apart from such analysis, researchers have also focused on visualization of Twitter data streams in order to assist the analyst to explore streams graphically.

In this study, some recent research work related to storage of the large volume Twitter data streams for performing analysis at large scale has been studied. At the end these approaches are critically evaluated in terms of purpose, data storage methods and technologies used.

LITERATURE REVIEW

Bringay *et al*. [7] focused on defining and manipulating the tweet cubes in a multi-dimensional architecture. The study proposes to create a STAR Schema in a multi-dimensional environment for analysis of data stream from Twitter. In the first step of the process, the authors define a STAR schema for the data warehouse by exploring the dimensions and facts from the tweets data. The second step in the process used TF-IDF Adaptative measure to find out more significant words in the hierarchy levels of the cubes. The model identifies the context of tweets by analyzing the word hierarchy using MeSH (Medical Subject Headings), a vocabulary for medicines, if no predefined hierarchy exists in the first step. Although the authors are able to classify the most important words in dimension hierarchies, the approach seems to work for only STAR Schema for multidimensional architectures. Moreover, the contextualization is done using Medical Science Library, and thus the hierarchal contextualization only works for medical sciences if a predefined hierarchy is not available.

A similar approach on text cube mining was presented by Liu *et al.* [8] in which they aimed to study the human social cultural behavior using Twitter stream data. For this purpose, the authors created text cube architectures in order to develop prediction models. The framework proposed in this research has three parts. The framework extracts linguistic features from the data in the generic language function. The next immediate part enables the selection of linguistic features from the list by exploiting the psychological and sociological expectations. The HSCB layer is added to evaluate the given dimensions. These dimensions and facts are used to create a STAR Scheme at the next stage. The prediction models are applied at the last step on the data warehouse for each dimension using the selected features.

The methodology is tested for prediction of violence in Egypt during 2011. However, the methodology is not tested with huge amount of data.

In another approach, Rehman *et al*. [9] introduced a technique which created a data warehouse based upon the Twitter streams. The architecture presented in the technique works on 5 layers. The first layer is used for capturing Twitter Streams using Twitter APIs (REST API, Search API or Streaming API). Second layer in the process reads the data available in JSON Format and creates an XML. MS SQL Server is used for saving the data in data warehouse architecture in the third layer. The remaining two layers in the architecture include front end tools for analysis and presentation purposes. The technique is tested on the data gathered from Twitter stream related to earthquake in Indonesia. Data was stored in Data warehouse, analyzed and was mapped graphically for visualization purposes. The authors intend to perform content analysis involving spam filtering and language detection from texts. Authors also suggested the future directions like usage of Yahoo, Wikipedia and Similar services for events and entity detection.

Extending their work further, Rehman *et al.* [10] worked on analysis of unstructured and semi-structured data from Twitter data streams. In the first step of this methodology, the authors extract facts and dimensions required for multi dimensional STAR Scheme to build the data warehouse. These facts and dimensions are extracted on the basis of aggregated information formed by semantic analysis. Furthermore, two APIs are used for semantic analysis including Alchmey API and Open Calais. These APIs are used for sentiment analysis and topic generation respectively. A sentiment analysis is performed which categories the tweets into Negative, Neutral, Positive and No Sentiment categories. Afterwards, entity and event detection is performed to get further insights according to the entities like Persons or Countries. Such Aggregate data is then stored in the data warehouse, where data mining techniques are applied to explore the aggregated information for analysis purposes.

Liu *et al.* [11] also extended their work further by studying different kinds of human, social and cultural behaviour (HSCB) embedded in the Twitter stream. The authors have extended their previous approach of sentiment analysis on text cubes by displaying contents of cubes over a heat map along with application of mining models on the data from these cubes. To start the process, linguistic features are extracted particularly emotional features. At this step, these features are treated as measures, whereas the sentiments are targeted as facts to create a STAR Schema for the data warehouse. Afterwards data mining methods like LIBSVM, REGTree and IBL methods are used to exploit the data warehouse for analysis purposes. The contents of cubes are further mapped on a heat map, where degree of opacity indicates the value of the behavioral, social or cultural measures, helping analysts to focus on concerned hotspots on the map. The methodology is tested with two data sets taken from a U.S. City, and Arab Spring.

In a similar approach for behavior detections, Vioulès *et al.* [21] presented a methodology to automatically identify a sudden change in a Twitter user online behavior. In this technique, the authors have combined natural language processing techniques with some text and behavioral features. The experiments show that this scoring method is able to capture the warning signs better than compared with some other machine learning algorithms like Random Forest, Simple Logistic and J48 Classifiers.

A location based study was performed by Kumar *et al.* [12], in which they worked on identification of a tweet whether it is coming from a crisis region or not. Authors argue that only 1% of the tweets include geographic information, and therefore it is not easy to confirm if the tweet is from a crisis region or not. Moreover, in such cases, it is a time consuming process to analyze the history of the tweeter to determine the same. Authors focus on the tweet text instead in order to identify as if it is coming from a crisis region. Authors have found that if a tweet is from crisis region, then its more likely to discuss a novel topic, it is less likely to seek attention and it is more likely to use external resources to convey their message. The authors have used different parameters like mobile features, resource features, textual features, linguistic features, and user features to identify behavioral patterns,. The experiments were conducted using Naïve Bayes and Random Forest Classifier through Weka Tool. In order to find out the most important feature among the given features, logistic regression was applied. It was found that linguistic features were the most important class of features. A case study performed on Arizona Wildfire Tweets by using the described approach, and it was found to be more effective than another similar approach. In future, the authors intend to extend this study in terms of size of training data in order to see its effect on classification.

In a different study, Koupaie *et al.* [13] presented an algorithm which is used to detect outliers in data streams by involving a clustering method. According to the authors, the unsupervised data mining approaches are more feasible when compared with the supervised approaches, because these don’t require class labels of objects. Moreover, such techniques can detect unforeseen outlying cases. Moreover, these techniques don’t require knowledge of data in advance. Authors presented a cluster based outlier detection algorithm for data streams, where real time outlier is detected in stream data using incremental clustering algorithm. The methodology presented in this paper has not been compared with any other similar approaches to draw a comparison for accuracy and efficiency.

Kraiem *et al*. [14] worked on OLAP Analysis of Twitter data stream, whereas a multidimensional model was proposed. The methodology works on specificity of the tweet text and creates a linkage between the tweets and its responses. The text of the tweets is studied to create a multidimensional architecture. A logical model is created by applying different set of rules. A tweet OLAP prototype (OLAP4Tweet) is created to analyze the Twitter accounts. This prototype is created in Java with Oracle Database. In the first step of analysis, the stream of data is read into the database for a couple of days. As the next step, the language detection is done by using a JSON-based API. The tweets are categorized per language at this stage. Afterwards the user category (Information Seeker, Friendship/Relationship, and Information Sharing) is used to allocate the data within these categories. In the text step, Tweet Type (Normal Tweet, Responses, Mentions, Retweet) is setup against each tweet by analyzing the data. OLAP Cubes are further created to perform the actual analysis as per requirements. The methodology allows Twitter data exploration; however it requires the analyst to setup predefined categories, tweet types and other dimensions, therefore, it is not a generalized approach which can be used for specific cases.

In order to test the effectiveness of open source databases, Murazza and Nurwidyantoro [15] tested Cassandra NoSQL Database to create a data warehouse using real time Twitter data. Authors have written data into the Cassandra NoSQL database, and performed read operations on the same. The results obtained were compared with the other relational databases like MySQL and PostgreSQL for both write and read operations. It was found that the Cassandra provides better write operation time, compared with other databases. However, it is not faster than other databases in terms of read operations. The authors have created visualization features using JQuery Libraries to show real time hash-tag counts. However, the real-time visualization requires a one second gap after every 20 seconds to load the new data. Authors have suggested that the research work can be enhanced in two directions. One of the directions is to cover the join and aggregation limitations, which can be done by integration it with some real-time analytic framework. Other future direction is to improve the performance by creating clusters of data instead of applying analysis on all data at once.

In order to receive Twitter data, perform analysis and provide visualization capabilities, Sechelea *et al.* [16] presented a methodology which exploits the geo-location parameters. Authors have created a script in Python which harvests the Twitter data based upon some geo-location parameters and saves data into a text file. Afterwards, the data is preprocessed by removing re-tweets, applying tokenization, removing stop-words, and applying stemming. Afterwards, clusters of data are created using a consensus matrix along K-Means Algorithm and DBSCAN algorithm. The clustering technique allowed the authors to obtain hot topics from the data which is collected and preprocessed in the previous step. Open Street Maps are used to provide the 3-D visualization capability through a python script. The visualization shows the density of Twitter activity in a given location geographically.

Bordogna *et al.* [17] proposed a new technique to exploit time-stamped geo-tagged tweets posted by the users. The tweets in this research work are tracked to trace their trips utilizing a clustering algorithm. The algorithm focuses on grouping similar tours together to analyze popular tours. The overall framework created by the authors is termed as The Interoperable Framework for Trips Tracking and Analysis. The framework includes two tools named as Follow Me Suite and Trips Analysis Suite, which are used for Trips Identification and Geographical Analysis respectively. Another component in the architecture is called as Geo Server which is used to publish the data of trips as open data for further analysis through OGC Standard Geo Portal Clients available online. The fourth component is the Tourist Tracker Portal that enables online analysis of popular tours. Authors intend to integrate this tool with other social network platforms. Moreover, in future it was aimed to utilize another clustering technique to exploit unexpected knowledge from gathered trips.

Cuzzocrea *et al.* [18] presented a new methodology to target the arrangement of Twitter tweets in OLAP Cube within a conceptual hierarchy. To achieve this, the authors have integrated Time Aware Fuzzy Formal Concept Analysis Theory with OLAP Technology over multidimensional tweet streams. Authors have also introduced a summarization algorithm which creates subjects of tweets according to the analysis perspectives.

In a recent research, Mallek *et al.* [19] presented an approach to create a data warehouse for Twitter data which uses NoSQL database for storage of data imported from Twitter. The tool created by the authors is termed as BigDimETL which processes the unstructured data received from Twitter streams into a NoSQL Database and then uses MapReduce functions authored by Google's Hadoop Framework to load the data into a data warehouse based upon STAR Schema. The data warehouse is further exploited to perform different kinds of analysis.

In another recent approach, Ibrahim *et al.* [20] conducted a study to discuss different techniques for topic detection from Twitter data stream. Authors have tested SFM, Bngram, CSS and Examplar-based topic detection techniques on 3 Twitter data sets and calculated their performance using term precision, term recall and topic recall along with the running time. It was found that SFM and Bngram provide better precision among the techniques. Moreover, CSS performs better in terms of topic and term recall. From the study, it can be concluded that each method has its advantages and disadvantages. The business analysts can select a certain technique by adapting the required precision and recall values. It will be interesting to see these techniques in terms of F1 measure which seeks a balance between both of these measures.

Summary of previous research

A brief Summary of Techniques used in mining of Twitter data stream is presented in Table 1. Furthermore, the technological summary of these techniques is given in Table 2. According to the findings, it was found that Twitter Streaming API was used in most of the studies to receive the data in JSON format and store it locally for processing purposes [2], [7], [8], [14], [17], [18]. However, some studies focus on geo-locations while receiving the data, resulting in location-specific analysis. Some approaches targeted specific cities instead of tweets from all over the world. At times, the streaming data was received using a specific event, like a crisis a political campaign or a mega event like a football final [12], [17]. Such studies show that business analysts from all domains have targeted the Twitter data stream for knowledge discovery.

There are 67 variables in a tweet [10]. Some of these are numeric, whereas some variables involve timestamps. One of the major attributes is tweet text itself. Most of the studies have focused on tweet text for semantic analysis, textual mining, opinion mining, topic exploration or behavioral analysis [11],[12], whereas other studies have utilized the associated Meta tags present with the tweets for knowledge discovery [9],[18].

It is also interesting to review the storage mechanism in different approaches used for knowledge discovery from Twitter data streams. Some approaches have preferred to use text files, whereas there are approaches where XML objects are used to store the extracted data [10]. The usage of open source databases like Cassandra, MySQL, PostgreSQL etc. has also been witnessed [15], [17]. The approaches which have targeted multi-dimensional analysis have preferred to use SQL Server to store the retrieved information. In multi-dimensional mining of data streams, most of the authors have used STAR Schema as the multi-dimensional architecture in order to store the aggregated information for analysis purposes [8], [11].

The targeted objectives in the reviewed research mainly include the following: topic identification, location based intensity of Twitter usage, improvements in storage time of data streams, application of OLAP over multidimensional stream data, behavioral analysis, crisis detection and other events detections. These objectives are either met by analyzing the tweet text or the analysis done using Meta information of tweets in a multi-dimensional environment [9], [14].

It is also found that most of the techniques include a visualization component to assist the analyst to explore the tweet streams graphically. Some of the researchers have created pie charts or column bar charts to display the aggregated data from the databases. Whereas some researchers have tried to plot the Twitter activities on Google Maps or similar components to display the location based intensity of Twitter data streams. Some of the researchers have created Heat Maps as well. For visualization purposes, researchers have utilized open source and licensed components including GeoServer, Google Map Services, Wed Map Services, HoloVizio and JQuery High Chart Components [15]-[17].

Different tools, technologies, algorithms and libraries have been used in the previous researches for such analysis which include open source and licensed databases/languages like MySQL, PostgreSQL, PHP, SQL Server, JQuery, JSON and XML.

Discussion and Future Recommendations

A discussion on the gaps identified in the previous researches is given below along with the future recommendations to fill these gaps:

1. Almost all approaches require data to be in file system, database or a data warehouse in order to process for analysis purposes. Some approaches are using XML structures; therefore these approaches take more time to process the data.

2. The approaches which are storing data to the database are lacking techniques or a mechanism to deal with continuous incoming data streams. In our opinion, there should be a precise way to continuously update the underline storage with the latest data stream so that a nearly real-time analysis is possible.

Table 1: Summary of Techniques used in mining of Twitter data stream

No.	Title	Objectives	Findings/Achievements	Limitations
1	Towards an on-line analysis of tweets processing - [7]	-Development of a data warehouse for multi-dimensional analysis -Contextual Analysis of Multidimensional data	-Usage of TF-IDF Adaptative Measure for finding significant words at hierarchical levels. -Usage of MeSH Library for Hierarchical levels in case of absence of predefined hierarchies	-The Approach has targeted dietary domain. It will be interesting to test the methodology in other domains.
2	SocialCube: A text cube framework for analyzing social media data - [8]	Multidimensional Modeling of Twitter Data Stream for Information Querying and Visualization	-Human, Social, Cultural and Behavior (HSCB) Analysis -Predictive HSCB Modeling	-Methodology works effectively for political events but can be explored further on variety of datasets. -HSCB Analysis is done in a textual database which can increase the computation time.
3	Building a Data Warehouse for Twitter Stream Exploration - [9]	To Demonstrate the power of Data warehouse technology for data stream analysis	-Knowledge Discovery using multi-dimensional schema for Twitter data stream	Relies heavily on topics already defined by Twitter, so hidden knowledge discovery is very limited.
4	OLAPing social media: the case of Twitter - [10]	Multidimensional Modeling of Twitter Data Stream for Text and Opinion Mining	-Semantic Analysis on multi-dimensional Twitter stream data -Event Detection	For opinion mining the targeted APIs have a limit per day, which can be a hurdle for the mining process.
5	A text cube approach to human, social and cultural behavior in the Twitter stream – [11]	Multidimensional Modeling and Visualization of Twitter Stream Data for HSCB Study	-Human Social and Culture Behavior Study -Text Cube Visualization on Heat Map	-Methodology works for political events effective and it can be explored to see its effectiveness on variety of datasets. -HSCB Analysis is done in a textual database which can increase the computation time.
6	A Behaviors Analysis Approach to Identifying Tweets from Crisis Regions - [12]	Identify a tweet as if it's coming from a crisis region	-Behavioral Analysis using Tweet Contents	-Requires rules to define the crisis behavior. -A visualization capability will suit the topic
7	Outlier Detection in Stream Data by Clustering Method – [13]	An algorithm to detect outliers in data streams	-Outlier detection in data streams using clustering -Accuracy better than other algorithms	-No experiment data is available to see the methodology in place -The comparison of other approaches is un known
8	Modeling and OLAPing social media: the case of Twitter – [14]	Multidimensional Modeling of Twitter Data Stream for Information Retrieval	Development of Constellation Schema for Multidimensional modeling of Twitter data	The exploration works effectively but techniques like topic exploration and automated schema generation can be handy if integrated.
9	Cassandra and SQL Database Comparison for Near Real-Time	Comparison of Cassandra and SQL Server for data storage and	-Cassandra has better storage ability for streaming data	-The analysis power of the approach is not known as the focus is on time analysis

No.	Title	Objectives	Findings/Achievements	Limitations
	Twitter Data Warehouse – [15]	multidimensional analysis for streaming data	-Cassandra has lower query performance for data retrieval	-It will be appropriate if the approach is compared with conventional OLAP and other data mining techniques
10	Twitter Data Clustering and Visualization – [16]	Identification of topics from Twitter data stream Detection of intensity of Twitter activity at a location	-Topic Clustering -Visualization of Tweets in Density and 3-D Maps	-No Pruning of topics is done, so garbage topics are expected. -Visualization remains overly technical for business analysts
11	Clustering Geo-Tagged Tweets for Advanced Big Data Analytics – [17]	-Group Similar Trips based upon Geo-Location to analyze popular tours	-Grouping of Tweets based up the geo-location -Visualization Ability of Geo-Tagged Tweets	-Highly focused on Geo-Location which is not publically available most of the time
12	OLAP Analysis of Multidimensional Tweet Streams for Supporting Advanced Analytics – [18]	-Integration of FFCA with OLAP -A microblog summarization algorithm to reduce the tweets data to a subset	-Integration of FFCA with OLAP. -Usage of Microblog Summarize for data reduction	-It is not known if the discarded data in the process has an impact on the analysis -A comparison with the full data could have helped as a comparative study
13	BigDimETL with NoSQL Database – [19]	Usage of NoSQL database for data warehouse for Twitter data stream	-A tool called BigDimETL having ability to create a NoSQL Database for data warehouse creation -Usage of Map-Reduce functions to load data into data warehouse	-A comparison with similar approach can be done to see the effectiveness of using NoSQL rather than traditional relational databases
14	Tools and Approaches for Topic Detection from Twitter Streams: Survey – [20]	Comparison of techniques for topic detection from Twitter Data Stream	-SFM and Bngram provide better precision -CSS performs better in terms of topic and term recall	-The approach requires technical knowledge of Matlab. -F1 Measure can be integrated in the system for the comparison purposes.
15	Detection of Suicide-Related Posts in Twitter Data Streams – [21]	Sudden Change Detection in user's online behavior	-Combined Natural Language Processing Methods with Textual and Behavioral features for user's behavior detection -Experiments conducted on datasets and compared with Machine Learning Classifiers	-The behavioral features are different for different individuals, so the approach can be improved by testing different behavioral features for different persons instead of fixed behaviors.

Table 2: Technological Summary of Techniques used in mining of Twitter data stream

Sr. No.	Title	Database Storage	Experiment Dataset / Size	Visualization	Technologies / APIs / Libraries / Algorithms
1	Towards an on-line analysis of tweets processing – [7]	PostgreSQL	1,801,310 Tweets	No Visualization Ability	MeSH Library, TF-IDF Adaptative Measure, PostgreSQL
2	SocialCube: A text cube framework for analyzing social media data – [8]	Data Warehouse (STAR Schema)	Egyptian Revolution 64,000 Tweets	Visualization through ABMiner	JSON, ABMiner, LibSVM, RepTree, IBK, OLAP
3	Building a Data Warehouse for Twitter Stream Exploration – [9]	Data Warehouse (X-DFM based Schema)	Indonesia Earthquake 2012 86000 Tweets	-Time Series Charts -Pie-Charts	Microsoft SQL Server Analysis Services
4	OLAPing social media: the case of Twitter – [10]	XML Based Database (BaseX)	Euro 2012 Final 0.5 Million Tweets	Basic Column and Pie Charts	Alchmey API, OpenCalais, XML, BaseX
5	A text cube approach to human, social and cultural behavior in the Twitter stream – [11]	Data Warehouse (STAR Schema)	(Washington D.C Data Set) ~0.5 million tweets Egypt Revolt –Unknown Size	Visualization using Heat Map	OLAP, Twitter Streaming API, XML, libSVM, REPTree, IBK
6	A Behaviors Analysis Approach to Identifying Tweets from Crisis Regions - [12]	Not used	Arizona Wildfires (Size Unknown)	No Visualization Ability	Naïve Bayes, Random Forest Classifier, Wilcoxon Signed Rank Test
7	Outlier Detection in Stream Data by Clustering Method - [13]	Not Known	Not Known	Not Known	Not Known
8	Modeling and OLAPing social media: the case of Twitter - [14]	OLAP (Constellation Schema)	72,000 Tweets	Custom Charting Component (Column, Pie Chart Graphs)	Java, Oracle, JSON

Sr. No.	Title	Database Storage	Experiment Dataset / Size	Visualization	Technologies / APIs / Libraries / Algorithms
9	Cassandra and SQL Database Comparison for Near Real-Time Twitter Data Warehouse - [15]	Cassandra MySQL PostgreSQL	10,000 Tweets	Hashtag Counts shown over time using JQuery High Charts Component	Cassandra, PHP, JQuery Charts, MySQL, PostgreSQL, JSON
10	Twitter Data Clustering and Visualization - [16]	Not Known	250,000 Tweets from London 15,000 Tweets from Brussels	Heat Map using Google Maps 3-D Map using HoloVizio	Python, Matlab, C++, Apache
11	Clustering Geo-Tagged Tweets for Advanced Big Data Analytics - [17]	PostgreSQL	9127 Tweets from Bergamo, Italy	Visualization on GeoServer as per OGC Standard via Wed Map Service	Not Known
12	OLAP Analysis of Multidimensional Tweet Streams for Supporting Advanced Analytics - [18]	Data Warehouse (DFM based Schema)	2015 Italian Election Campaign (Unknown Data size)	No Visualization	Not Known
13	BigDimETL with NoSQL Database – [19]	NoSQL Database	Not clear	No Visualization	NoSQL, Hadoop, Map Reduce, XML, JSON, CSV
14	Tools and Approaches for Topic Detection from Twitter Streams: Survey – [20]	Not Known	FA Cup (74,00 Tweets) US Elections (620,000 Tweets) Super Tuesday (230,000 Tweets)	Matrix Visualization using TMG	Text to Matrix Generator (TMG) Matlab tool
15	Detection of Suicide-Related Posts in Twitter Data Streams – [21]	Not Mentioned	1: 5446 Tweets 2: 11,000 Tweets	Small Graphs	Not Known

3. Some approaches neglect the Meta tags while saving the data stream in the underline storage being used. The Meta tags have very detailed related information of tweet text, for example the location, timestamp, number of retweets and likes etc. which can be considered for a better analysis.

4. The approaches, which involve data warehouse for storage of aggregate data, don't have a mechanism to decide the dimensions as well as facts at run time. Such pruning may result in improving the storage capacity. It will further improve the analysis computation time. Moreover, a more targeted and meaningful analysis will be performed.

5. It can be seen that all these techniques are missing a one-window analysis framework. Different technologies and components are being used in isolation for performing certain operations. For example the data reading component is separate of the analysis component in most of the approaches. Moreover, the visualization component works separately from analysis component in some of the researches. A one-window framework will not only make the operations smooth, but will also reduce the computation time.

CONCLUSION

In this research study, the literature related to the mining of Twitter data stream has been studied. It is found that approaches using file systems to store the data streams have greater computation time. Moreover, there is a need to create a real-time data retrieval process for real-time analysis. Some approaches don't make use of Meta tags associated with Tweets, and the approaches which use such information don't prune these tags to ensure that only required tags are used for storage and analysis to avoid storage and computation issues. Moreover, there is a need to create a one-window analysis framework where all operations like data retrieval, storage, analysis and visualization are possible instead of working in different components in isolation.

REFERENCES

[1] X. Wu *et al.*, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, 2014.

[2] O. Maimon and L. Rokach, "Introduction to Knowledge Discovery and Data Mining," in *Data Mining and Knowledge Discovery Handbook*, Boston, MA: Springer US, 2009, pp. 1–15.

[3] M. Usman, "Multi-level mining of association rules from warehouse schema," *Kuwait J. Sci.*, vol. 44, no. 1, 2017.

[4] M. Usman and M. Usman, "Multi-Level Mining and Visualization of Informative Association Rules," *J. Inf. Sci. Eng.*, vol. 32, no. 4, pp. 1061–1078, 2016.

[5] M. Usman and M. Usman, *Predictive Analysis on Large Data for Actionable Knowledge: Emerging Research and Opportunities: Emerging Research and Opportunities*. IGI Global, 2018.

[6] A. Goritz *et al.*, *Analyzing Twitter Data: Advantages and Challenges in the Study of UN Climate Negotiations*. SAGE Publications Ltd, 2019.

[7] S. Bringay *et al.*, "Towards an on-line analysis of tweets processing," in *Int. Conf. Database and Expert Systems Applications*, 2011, pp. 154–161.

[8] X. Liu *et al.*, "SocialCube: A text cube framework for analyzing social media data," in *Int. Conf. Social Informatics*, 2012, pp. 252–259.

[9] N. U. Rehman *et al.*, "Building a Data Warehouse for Twitter Stream Exploration," in *IEEE/ACM Int. Conf. Advances in Social Networks Analysis and Mining*, 2012, pp. 1341–1348.

[10] N. U. Rehman *et al.*, "OLAPing social media: The case of Twitter," *IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. (ASONAM 2013)*, pp. 1139–1146, 2013.

[11] X. Liu *et al.*, "A text cube approach to human, social and cultural behavior in the Twitter stream," in *Int. Conf. Social Computing, Behavioral-Cultural Modeling, and Prediction*, 2013, pp. 321–330.

[12] S. Kumar *et al.*, "A behavior analytics approach to identifying tweets from crisis regions," in *Proc. 25th ACM Conf. Hypertext and Social Media*, 2014, pp. 255–260.

[13] H. M. Koupaie *et al.*, "Outlier Detection in Stream Data by Clustering Method," *Int. J. Adv. Comput. Sci. Inf. Technol.*, vol. 2, no. 3, pp. 25–34, 2013.

[14] M. Ben Kraiem *et al.*, "Modeling and OLAPing social media: the case of Twitter," *Soc. Netw. Anal. Min.*, vol. 5, no. 1, p. 47, 2015.

[15] M. R. Murazza and A. Nurwidyantoro, "Cassandra and SQL database comparison for near real-time Twitter data warehouse," in *Int. Seminar Intelligent Technology and Its Applications (ISITIA)*, 2016, pp. 195–200.

[16] A. Sechelea *et al.*, "Twitter data clustering and visualization," in *23rd Int. Con. Telecommunications (ICT)*, 2016, pp. 1–5.

[17] G. Bordogna *et al.*, “Clustering geo-tagged tweets for advanced big data analytics,” in *IEEE Int. Congr. Big Data (BigData Congress)*, 2016, pp. 42–51.

[18] A. Cuzzocrea *et al.*, “OLAP analysis of multidimensional tweet streams for supporting advanced analytics,” in *Proc. 31st Annual ACM Symposium Applied Computing*, 2016, pp. 992–999.

[19] H. Mallek *et al.*, “BigDimETL with NoSQL Database,” *Procedia Comput. Sci.*, vol. 126, pp. 798–807, Jan. 2018.

[20] R. Ibrahim *et al.*, “Tools and approaches for topic detection from Twitter streams: survey,” *Knowl. Inf. Syst.*, vol. 54, no. 3, pp. 511–539, 2018.

[21] M. J. Vioulès *et al.*, “Detection of suicide-related posts in Twitter data streams,” *IBM J. Res. Dev.*, vol. 62, no. 1, p. 7: 1--7: 12, 2018.