CHRIST
UNIVERSITY
BENGALURU, INDIA

Declared as Deemed to be University under Section 3 of UGC Act 1956

Dissertation Report on

Social Media Analytics on Twitter Content

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF TECHNOLOGY

in

Computer Science and Engineering

by

Nilakshi Roy 1427111

Under the Guidance of

Boppuru Rudra Prathap

and

Vijayavani Rao

Department of Computer Science and Engineering

Faculty of Engineering, Christ University,
Kumbalagudu, Bengaluru - 560 074

March-2016

CHRIST
UNIVERSITY
BENGALURU, INDIA

Declared as Deemed to be University under Section 3 of UGC Act 1956

Faculty of Engineering

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that **Nilakshi Roy** has successfully completed the dissertation work entitled "**Social Media Analytics on Twitter Content**" in partial fulfillment for the award of **Masters of Technology** in **Computer Science and Engineering** during the year **2015-2016**.

Boppuru Rudra Prathap
Guide

Vijayavani Rao
Co-Guide

Dr. K. Balachandran
Head of the Department

Dr. Iven Jose
Associate Dean

CHRIST
UNIVERSITY
BENGALURU, INDIA

Declared as Deemed to be University under Section 3 of UGC Act 1956

Faculty of Engineering

Department of Computer Science and Engineering

BONAFIDE CERTIFICATE

It is to certify that this dissertation titled "Social Media Analytics on Twitter Content" is the bonafide work of

Name	Register Number
Nilakshi Roy	1427111

Examiners [Name and Signature]

1.

2.

Name of the Candidate :

Register Number :

Date of Examination :

ABIBA

March 29th, 2016

Ms.Nilakshi Roy

Christ University Faculty of Engineering

Bangalore

Subject: Certificate of Internship

Dear Nilakshi,

We Congratulate you in successfully completing your internship project of 4 months "Social Media Analytics on Twitter Content ",With ABIBA Systems Pvt Ltd. We wish you all the best in your future endeavours.

Yours sincerely,

For ABIBA Systmes Pvt Ltd.

ABIBA Systems Private Limited Bangalore

Poornima Shetty

Assistant Manager-HR

Acknowledgement

I would like to thank Christ University Vice Chancellor, **Dr. Rev. Fr. Thomas C Mathew**, Pro Vice Chancellor, **Dr. Rev. Fr. Abraham**, Director of Faculty of Engineering, **Fr. Benny Thomas** and the Associate Dean **Dr. Iven Jose** for their kind patronage.

I would like to express my sincere gratitude and appreciation to the Head of the Department of Department of Computer Science and Engineering, Faculty of Engineering **Dr. K. Balachandran**, for giving me this opportunity to take up this dissertation work.

I am extremely grateful to my guide, **Boppuru Rudra Prathap**, who has supported and helped to carry out the dissertation work. His constant monitoring and encouragement helped me keep up to the dissertation schedule.

I am extremely grateful to my co-guide, **Vijayavani Rao**, who has supported and helped to carry out the dissertation work. Her constant monitoring and encouragement helped me keep up to the dissertation schedule.

I would like to extend my sincere thanks to all the faculty members of Computer Science and Engineering department and the system administrator for providing logistics support. I also like to extend thanks to my friends and family members for their continuous support.

Declaration

I, hereby declare that the Dissertation titled "**Social Media Analytics on Twitter Content**" is a record of original dissertation work undertaken by me for the award of the degree of **Masters of Technology** in **Computer Science and Engineering**. I have completed this study under the supervision of **Boppuru Rudra Prathap**, Computer Science and Engineering and **Vijayavani Rao**, Analytics.

I also declare that this dissertation report has not been submitted for the award of any degree, diploma, associate ship, fellowship or other title anywhere else. It has not been sent for any publication or presentation purpose.

Place: Faculty of Engineering, Christ University, Bengaluru
Date: 26-03-2016

Name	Register Number	Signature
Nilakshi Roy	1427111	

Abstract

Daily millions of people all around the world use social media sites and twitter is the popular one. It has been changed the way people find information, communicate, and share knowledge and thoughts like good and bad business experiences, liking disliking about any company and opinions about the product. Social media analytics explains the way of measuring, analyzing and interpreting the outcome of interactions and associations among the people, topics and ideas. The most efficient use of social media analytics is to mine sentiments of the customer to support marketing and customer service activities. Twitter tweets for business and research activity is available due to the presence of web-based application programming interfaces (API) which allows to fetch tweet stream dynamically. In this project I have considered different network providers and fetched tweets of different users which have given some views about them. Twitter data warehouse is where along with the tweets, Meta data of twitter is stored in postgres database. Sentiment analysis, is the linguistics analytical feature on the tweets is made, where each tweet is labelled as positive, negative or neutral and stored back in database. This data base later store the data in facts and dimensions, known as data mart, a subset of the data warehouse. Following data mart helps to create an OLAP cube which is a better representation and visualization of data. Another approach of representing the data from the data mart is predictive analysis, which listens the history of the data and forecast the upcoming data over a period of time. These two approaches are used by the business users for creating reports and also for taking precautions according the changes in the data.

Contents

CERTIFICATE i

BONAFIDE CERTIFICATE ii

INDUSTRY CERTIFICATE ii

ACKNOWLEDGEMENT iv

DECLARATION v

ABSTRACT vi

LIST OF FIGURES viii

LIST OF TABLES ix

GLOSSARY x

1 INTRODUCTION 1
- 1.1 Problem Formulation 5
- 1.2 Problem Identification 5
- 1.3 Problem Statement & Objectives 6
- 1.4 Limitations 6

2 RESEARCH METHODOLOGY 7

3 LITERATURE SURVEY AND REVIEW 10
- 3.1 Literature Collection & Segregation 10
- 3.2 Critical Review of Literature 18

4 SOCIAL MEDIA ANALYTICS 23
- 4.1 Methodology for the Study 23
- 4.2 Analytical Work 35
- 4.3 Design 38

5 RESULTS AND DISCUSSIONS **42**
5.1 Results & Discussions 42
5.2 Conclusions 57
5.3 Scope for Future Work 58

BIBLIOGRAPHY **59**

PUBLICATION DETAILS **61**

A Appendix Source Code **62**

B Appendix Paper Publication **66**

LIST OF FIGURES

1.1 components of data warehouse 5

3.1 Relational View of Tweet Record 16
3.2 Sentiment Analysis Methodology 16
3.3 Steps for Data Analytics 20

4.1 Work Flow Diagram 23
4.2 How twitter API works 25
4.3 ETL process 27
4.4 Using Naive Bayes Algo sentiment analysis 29
4.5 Twitter Data Warehouse 38
4.6 Star Schema model with facts and dimension table 41

5.1 Front page of my application. 42
5.2 For all dates the measures like length and score is calculated 43
5.3 Date and sentiments dimension with score measure 44
5.4 Graphical view of Fig 5.3 44
5.5 Sentiments with length and score measure 45
5.6 Sentiments with dates dimension and scores as a measure 45
5.7 Dates will levels calculating score and length 46
5.8 Sentiments with time dimensions calculating measures are length and score 47
5.9 All users and network provider where measures ae following count, followers count 47
5.10 All the network provider with measures sentiments,following count, followers count 48
5.11 Graphical form of Fig:5.10 48
5.12 network providers, sentiments and time dimension, following count and followers count measure 49
5.13 All dates and network provider, following count and followers count as the measures 49
5.14 All network providers,all the time, with measure following count and followers count 50
5.15 Sentiments and month name as in dimension and tweet count,following count and followers count in measures 50
5.16 Graphical representation of FIG 15.5 51
5.17 All network providers and sentiments with measure tweet count 51

5.18 Graphical representation of FIG 15.7 51
5.19 All the network providers trend value for 3 continuous month taking measures like score 52
5.20 All the network providers trend value taking measures like following count and followers count 52
5.21 All the network providers trend value taking measures like tweet count . 53
5.22 All the network providers and the sentiments, trend value for 3 continous month taking measures like following count and followers count . 53
5.23 Network provider dimension with retweet count and score measures . . 54
5.24 Predicted trend for all network provider with the measures 54
5.25 Predicted trend for all network Provider with the measures 55
5.26 Predicted trend for all network Provider and sentiments with the measures 55
5.27 Predicted trend for all network Provider with the measures 56
5.28 Predicted trend for all network Provider with the measures 56

LIST OF TABLES

4.1 Cube view data I 39
4.2 Cube view data II 39
4.3 Cube view data III 39
4.4 Cube view data IV 39
4.5 Time Dimension 39
4.6 Date Dimension I 40
4.7 Date Dimension II 40
4.8 Network Provider Dimension 40
4.9 Sentiment Dimension 40
4.10 User Dimension 40
4.11 Tweet Dimension 40
4.12 Tweet Fact 40
4.13 User Fact I 40

GLOSSARY

.

Item	Description
OLAP	Online Analytical Processing
API	Application Program Interface
ETL	Extract, Transform and Load
GUI	Graphical User Interface
NLP	Natural Language Processing
NLTK	Natural Language ToolKit
MDX	Multidimensional Expression
JDBC	Java Database Connectivity
SQL	Structured Query Language
UML	Unified Modeling Language
HSCB	Human Social Cultural Behavior
SVM	Support Vector Machine
SMA	Social Media Analytics

Chapter 1

INTRODUCTION

Outburst of social network activity in the latest years has led to generation of huge volume of user linked data which in turn gave the birth of the now a day's hot topic Social Media Analysis. Social media analytics determines the business goals of the data which is gathered and analyzed. Social analytics has an innovative means where companies can analyze their brand image among the customers. Top brand building marketing can be done in two ways one is traditional way where advertising, direct mail or many other one direction ways takes place and other is social media marketing. Social media marketing considered to be more difficult as messages spread to individual in a non- linear way.

Data warehousing is a technique for data warehouse which has become a good platform for most of the large companies. Companies, organizations and institutions all around the world have the opportunity to analyze public tweet stream to improve their marketing, customer services and also public relations from the learned knowledge. The increasing growth of web based social networks and the regular social activities results in huge social network data. Data can be of two type's one structured or unstructured like tweet, events, description, messages, brands, status, and comments. Data which has been stored in data warehouse as data cube must contain data of business process, customer reviews about some brand, distribution, sales, marketing etc. This data shows the customer pattern and trends, sentiments, business strategies and also many other characteristics. Data cube allows to aggregate numeric data and it is derived by measures and dimensions. So, the data which has been captured is valuable to the success of business and as well as to understand the human social behavior.

This work is to create a trend view for the public tweet stream for the purpose of its complete analysis. Twitter API allows the applications to get the dynamic data of tweet objects. Using API [4] automatic live tweets for a particular topic can be fetched and then transform the data into structured one later load them into particular database to do some successive analysis. In this paper we will show the analysis of social media can be useful from this established and developed technology. A trend view, which predicts the future that is a graphical form where the change of tweet count, retweet count, following count can be shown in a bar chart based on different time of different the network providers. We have introduced an OLAP cube approach for social media analysis, particularly sentiment analysis. The capabilities of OLAP has been extended by enrichment of the data set to find out new measures and dimensions for new data cubes and also supporting updated data as well as historical data. In data warehousing, data cube organize the data in multiple dimensions and several hierarchies for relevant data storing and visualizing from different perspectives. Data dimensional model, such as star schema, has been created. A star schema model has some proper dimensions and a fact table. As facts contain the events and the dimensions holds the reference information about fact. This schema is used to support data warehouse and data marts OLAP cubes.

1.0.1. Why twitter?

Twitter is the popular micro blogging network which supports real time information. Originally it's been presented in 2006 as a platform for interacting by sending short messages through internet. In recent years it has gain the worldwide popularity and also got into broadcasting news such as an influential channel and also the means of exchanging real time information. Around 332 million active twitter-users, over 500 million of tweets generates regularly till May 2015. Twitter has the attention of political, commercial, research and other establishments by allowing tweet stream available to the public. Twitter provides a set of APIs for fetching the detail information about the users and their communications and also the twitter streaming API for data-intensive applications, the search for API for querying and clarifying the message content and the REST API for retrieving the principal primitives of twitter.

1.0.2. Linguistic analysis for sentiment

Now a days the research are concerned on how sentiments are expressed in online reviews and news articles. This is a social and cultural way for human to express their emotions. Emotions reflect how an individual sees the world and also how he reacts to

the world in a particular event. Year's back cognitive psychologists didn't take sentiments of human in considerations. Understanding individual or analyzing the behavior of a group of people helps to predict the future and also to get the decisions for any business process. [17]Emotion can be assumed not only non-verbally but also using text-based communication. In recent studies it has been found that human convey their emotions using text, blogs, emails, short messages and different textual conversation. If a person is already sad before any start of conversation, the other participant can understand his negative emotion in text based communication. Emotions can be classified as negative positive or neutral.

Suppose you have used some particular product and you wrote some reviews showing your emotions about that product brand in social websites. Company fetches all those reviews of the customers about its product in some particular area and do proper sentiment analysis which helps to take proper business decisions.

Naïve Bayes classifier is a simple probabilistic model which is based on the Bayes rule with the strong independence assumption. This Naïve Bayes model gives a conditional independence assumption which has the class positive and negative bag of words which are conditionally independent to each other. This type of assumption doesn't harm the accuracy of the text classification but speed up the classification. If we use simplifying conditional independence assumption then the given class that is positive or negative words will conditionally independent of each other.

1.0.3. Natural language processing(NLP)

It is a arena of computer science, artificial intelligence and linguistics which includes the interactions between computers and natural languages. Specifically, it is the process of a computer which extracts meaningful information from natural language input and produces natural language output.

1.0.4. News analytics

The measurement of the qualitative and quantitative features of unstructured text data news tweets. Features like sentiments.

1.0.5. Star Schema

In data warehousing, star schema is the simplest form of any dimensional model, which has one or more fact tables which reference the dimension tables in order to structure the data. Fact table resides in the middle and it has surrounded by dimension tables.

It gives a structure like a star. Each dimension is characterized in a single table and the primary key for each dimension is linked to a foreign key in the fact table. The dimensions in the star schema is related to all the measures of the fact table.

1.0.6. Data Warehousing and OLAP

Data warehousing and Online Analytical Processing useful in different components of systems which helps in decisions in support and business intelligence. It has been originated in the 90s where they have granted the permission for giving the access to the data to decision makers. Components that consist of for building these type of systems are databases and applications that offer the tools analysts which needs to support the decision making in the organizations. In an organization data is the more important assets, these assets can be stored in two ways one is operational system of records and another is data warehouse. Operational system records store the data about taking orders, signing up new customers and logging complaints. It can query one row at a time which is quite time consuming and not Efficient. On the other hand Data warehouse store the data so that it can get the data from it and do operations. It basically deals with the count of new orders, why new customer signed, and why they have logged a complaint etc. It can access hundreds and thousands of row at a time. Data ware house is a specialization of database technology for integrating, accumulating, analyzing and also visualizing of data from different sources. It employed the multidimensional data model, which represents the data in a cube model which contains measures of interest. Data warehouse contains data that characterizes the business history of any company or organization. This ancient data is used for not only analysis; it also gives the provision of taking the business decisions at many levels something like strategic planning to routine evaluation of a discrete organizational unit. Fig1 shows the architecture of a data warehouse.

If the data in data warehouse represents multidimensional database then data are stored in OLAP cube. OLAP tools give the benefits of creating online statistical reports by the means of query and the analysis of data warehouse information. Summaries are calculated using aggregate functions like AVG, SUM, MAX, MIN many more. There are such cube operations like slicing, dicing, drill-up, drill-down, roll-up, pivot etc. is used by user to explore multidimensional cubes. Slicing is a procedure of selecting a subset of cube and taking a single value from the dimensions and also making new cubes from the rest of the dimensions. Dicing is an operation which creates a sub cube by permitting the analyst to pick some real values from the multiple dimensions. Drill-up and Drill-down where data can be concise and stretched from different ranges according to

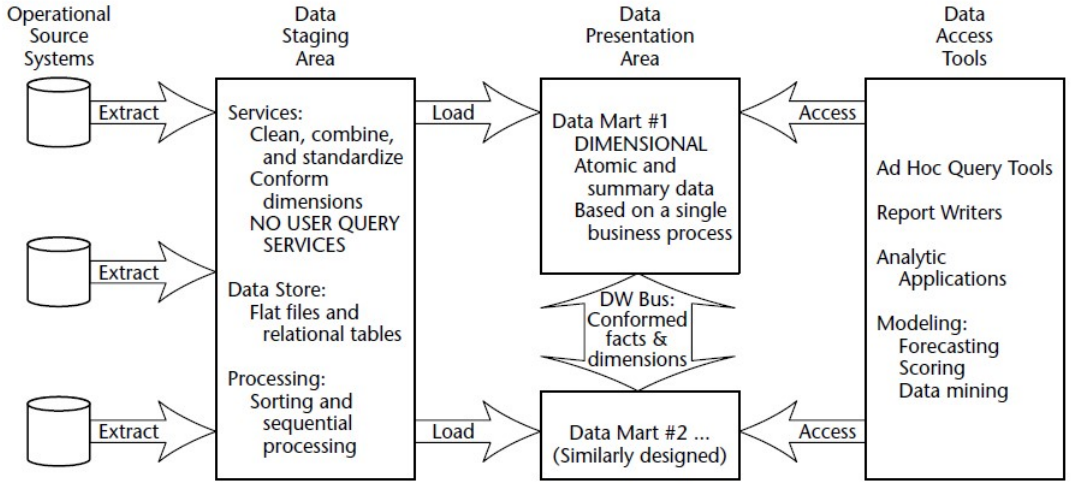


FIGURE 1.1: components of data warehouse

the user's choice. A roll-up operation which involves in briefing the data, along with the dimensions by using some formula. Lastly Pivot which allows the cube to rotate in space to see its different faces. Applicability of data warehousing is constrained to some business scenarios. Data analysis has become crucial in variety of applications like non-business domains such as government, science, education, hospital, research, medicine etc.

1.1 Problem Formulation

Social analytics listens and analyzes customer opinions. It analyze the sentiments of the customer which helps in business identifying customer satisfaction or any companies to analyze campaign that can be well received by the customers. It is therefore, behaving like a channel between customer and the companies, helping them to measure their influence in the market and also predicting the future of the organization.

1.2 Problem Identification

1) Representation of data in a multidimensional cube view for making business reports.

2) Predictive analysis such as identifying and predicting future outcomes and behaviors built on historical data collected over time.

1.3 Problem Statement & Objectives

Visualization and predictive analysis on multidimensional cubes. OLAP cube needs to be formed where we can visualize huge amount of data and make business reports, another is trend view is where we can do prediction about the future by observing the previous 3 months values.

1.4 Limitations

I have taken 1000 data from twitter for my project. This data has been collected since 2015-01-01. As I am doing predictive analysis, historical data is what helps to forecast the future. As much strong and amount of historical data is there in your data warehouse, the accuracy of predicting the future will be more specific. I have limited amount of data which may not contains more information about the history, so the accuracy may differ.

Sentiment analysis has been done using Naïve Bayes classifier. This classifier has the training set that is positive and negative. This two training set trains each word in the document in test data and trains it. Based on your training set, the accuracy of the positive and the negative sentiments comes. This training set is manually built as predefined twitter training set data is not available. Thus the accuracy differs.

Chapter 2

RESEARCH METHODOLOGY

Research refers to the search of new and useful knowledge based on a particular topic. It is a logical and systematic search for information. Many people have different views on research definition. They have given a brief explanations. Among many researchers, one has proposed that as “a careful investigation or survey for the new facts in any branch of knowledge. Research can be thought of moments towards the unknown from the known. Research is a knowledge which discovers the hidden truth of any topic. Information basically relates about matters. The information is collected from various sources a like experience, books, journals, human beings, nature, social media, television etc. A research can be made to move when there is a huge contributions of the knowledge in the current information. Only through research you can proceed and make progress in a field. Research is not always linked to science and technology, it can be on a vast area in other disciplines such as languages, literature, sociology and history. Research may differ from different topics but the research should be active, diligent and systematic process. Research is basically to learn more in depth to some topic, discover, interpret and revise the facts, events and the behavior or theories. Experiments, observations, analysis, comparison and reasoning is the way the research has to be done.

The prime objectives of research are to invent new facts on a particular topic on which you are doing research, verifying and testing the important facts. Analyze the event and the situation and to understand the cause and the effects, Also developing new tools, theories, algorithm and concepts are the part of the research which finds and solves new problem scientifically , nonscientific ally. Research has importance both in scientific and non-scientific fields. In our life problems occur daily, some problems can be solved using predefined solutions given or the algorithm and theories and some problems can

be solved by creating new solutions concepts and algorithm. Indirectly research make us to understand the nature and the natural phenomenon. A research problem can be the difficulty in a scientific community, government organization, anyone's social life and industry. Research on the predefined theories and solutions helps us to find out the wide range of the applications on them. It gives huge amount of knowledge and the proper way for solving problems. Research is also important in the industry and the business where problem involves in increasing the quality of products which leads to the understanding of the new materials, product, the way of living life and the new stars. In short researchers gives us a life of new style and makes it delightful and glorious. Research methodology is the way to solve a problem in a systematic way. It is a study of science where it has been explained how a research can be carried out. It is a way by which a researchers describes their work, explains it and predicts the occurrence. It can be followed as a work plan through which we study, gain knowledge and solves problem.

In social media is the place where users interact with each other, share their digital content. Daily huge amount of data has been generated. Twitter is one of the famous social microblogging sites where people shares their views with 140 characters. Social media analytics gathers the data from the social media, as here is twitter and analyze it.

Every users has their views on a particular topic. Suppose when a company launches their product, many users must have tweeted about the product, reviews can be bad or good. Data can be collected from the twitter by the help of twitter API. This twitter API helps to collect tweets based on any topic. The json file is cleaned and stored in database in a structured form. Data which is stored in database as a data mart where data is represented in facts and dimensions. For better analyzing, the linguistic analysis, such as sentiment analysis has done. This is the process where a business users can understand the real emotion of the user on that particular topic. Sentiments can be of positive, negative and neutral. Sentiment analysis now a days has taken a big part of analyzing and taking proper business decisions. For sentiment analysis Natural language Tool Kit has been used. Naïve Bayes classifier is one of the NLTK that used for doing sentiment analysis. As it has training set classifier which trains the test set data. There are two training set one positive and another is negative set data. For better understanding of data, data is represented in facts and dimensions. This creates multidimensional data model name OLAP cube.

Trend view is a representation of data where the changes in measures over certain amount of time can be visualized by different colors. Suppose for a company name,

orange, it has launched a product, where in the first month users followed the advertisement tweet and also retweeted. And in the next two months gradually they didn't response much. Since, due to lacking of interest the measures will become less. For each 2 percent of changes in data, the color will change. And also it views the last three months data. This trend view is also created using facts and dimensions table in the PostgreSQL. Mondrian is an open source Online Analytical Processing server in java , which responds to the query fast enough to allow an interactive exploration of data , even there are millions of records, occupying gigabytes of data. Tomcat is an application server from the Apache software foundation that implements Java servlets and Web pages that include Java server Page coding. Tomcat helps to show the results of the open collaboration of the developers. Mondrian use OLAP4j API which is an extension of JDBC for OLAP applications that use XML provides for analysis. The cube schema is written in xml file where we mention the dimensions and the level of hierarchy of the dimensions we want to keep .and the measures along with the operations like sum, count etc. we want to display. This helps to create an OLAP cube. When we start Tomcat application, the Mondrian automatically generates MDX query that is multidimensional expression query. This MDX query helps to plot and to display the data in dimensions and measure.

Chapter 3

LITERATURE SURVEY AND REVIEW

3.1 Literature Collection & Segregation

Researchers [1]are concerned on how sentiments are expressed in online reviews and news articles, blogs or social media. This is a social and cultural way for human to express their emotions. Emotions reflect how an individual sees the world and also how he reacts to the world in a particular event. Year's back cognitive psychologists didn't take sentiments of human in considerations. Understanding individual or analyzing the behavior of a group of people helps to predict the future and also to get the decisions for any business process. Emotion can be assumed not only non-verbally but also using text-based communication. In recent studies it has been found that human convey their emotions using text, blogs, emails, short messages and different textual conversation. If a person is already sad before any start of conversation, the other participant can understand his negative emotion in text based communication. Emotions can be classified as negative positive or neutral. Suppose you have used some particular product and you wrote some reviews showing your emotions about that product brand in social websites. Company fetches all those reviews of the customers about its product in some particular area and do proper sentiment analysis which helps to take proper business decisions. Naïve Bayes classifier is a simple probabilistic model which is based on the Bayes rule with the strong independence assumption. This Naïve Bayes model gives a conditional independence assumption which has the class positive and negative bag of words which are conditionally independent to each other. This type of assumption doesn't harm the

accuracy of the text classification but speed up the classification. But here author has explained that Naïve Bayes classifier is not enough to get the accurate measure. So he has implemented a combination of two three methods like effective negation handling, word n-grams and feature selection for the correct and proper accuracy. This is the fast sentiment classifier which gives the accuracy of 88.08 percent. Author has done linear training on IMDB movie reviews. Negation handling is if a word comes in text nice , it is considered as positive word but if it comes as not nice, still as this nice is a positive word, it will comes as positive, not will be recognized. He has taken an approach to solve this problem in sentiment analysis. N-grams is parting the text into bigrams, trigrams and unigrams. When data is fetched, lots of redundant values come. To remove the redundant value, feature selection can be done which removes all the disambiguation capabilities. Author has user Bernoulli Naïve Bayes Classifier, where each word counting is done once from the document. It gives overall 88.80 percent accuracy of the 25000 of test set data.

Social media analytics[2], SMA has evolved a lot in recent years. In this appear author has mainly focused on the customer both external and internal organization environment. Business SMA had some benefits has helps the farm to get a potential growth and also understanding the customer need and the problem they are facing. First is improving marketing strategy that is customer reviews and information about the product they have used it posted on blogs, papers, news, and social media. For market analysis these information's are used and preprocessed in a structured way and so that company can make good market decisions. Another is for better customer engagement it has monitored the best customers which used to purchase more or often purchase and posts. According to their views company make business decisions and make effective precautions. For better customer service it has observed they needs of the customer, and SMA sees how a customer faces the problem. This can be resolved to satisfy more customers. Reputation of a company is the main basic thing for success. If the company does not satisfy its customers, it will get bad reviews. Through social media, blogs, news everywhere, people will be discussing and thus the reputation goes away. Business decision makers should focus on their happening customer to satisfy and keep their reputation string.

Researchers[3] are concerned on how sentiments are expressed in online reviews and news articles, blogs or social media.. This is a social and cultural way for human to express their emotions. Emotions reflect how an individual sees the world and also how he reacts to the world in a particular event. Year's back cognitive psychologists didn't

take sentiments of human in considerations. Understanding individual or analyzing the behavior of a group of people helps to predict the future and also to get the decisions for any business process. Emotion can be assumed not only non-verbally but also using text-based communication. In recent studies it has been found that human convey their emotions using text, blogs, emails, short messages and different textual conversation. If a person is already sad before any start of conversation, the other participant can understand his negative emotion in text based communication. Emotions can be classified as negative positive or neutral. In this paper author has introduced POS specific for priority feature and explore the tree kernel. Suppose there is a statement which is represented in a tree structure. Initialize with the root node, then each time you tokenize each tweet. If the token is target, negation, punctuation etc. it will be tagged in the tree structure.

Data warehousing and Online Analytical Processing[4] useful in different components of systems which helps in decisions in support and business intelligence. It has been originated in the 90s where they have granted the permission for giving the access to the data to decision makers. Components that consist of for building these type of systems are databases and applications that offer the tools analysts which needs to support the decision making in the organizations. In an organization data is the more important assets, these assets can be stored in two ways one is operational system of records and another is data warehouse. Operational system records store the data about taking orders, signing up new customers and logging complaints. It can query one row at a time which is quite time consuming and not Efficient. On the other hand Data warehouse store the data so that it can get the data from it and do operations. It basically deals with the count of new orders, why new customer signed, and why they have logged a complaint etc. It can access hundreds and thousands of row at a time. Data ware house is a specialization of database technology for integrating, accumulating, analyzing and also visualizing of data from different sources. It employed the multidimensional data model, which represents the data in a cube model which contains measures of interest. Data warehouse contains data that characterizes the business history of any company or organization. This ancient data is used for not only analysis; it also gives the provision of taking the business decisions at many levels something like strategic planning to routine evaluation of a discrete organizational unit.

If the data[5] in data warehouse represents multidimensional database then data are stored in OLAP cube. OLAP tools give the benefits of creating online statistical reports by the means of query and the analysis of data warehouse information. Summaries are calculated using aggregate functions like AVG, SUM, MAX, MIN many more. There

are such cube operations like slicing, dicing, drill-up, drill-down, roll-up, pivot etc. is used by user to explore multidimensional cubes. Slicing is a procedure of selecting a subset of cube and taking a single value from the dimensions and also making new cubes from the rest of the dimensions. Dicing is an operation which creates a sub cube by permitting the analyst to pick some real values from the multiple dimensions. Drill-up and Drill-down where data can be concise and stretched from different ranges according to the user's choice. A roll-up operation which involves in briefing the data, along with the dimensions by using some formula. Lastly Pivot which allows the cube to rotate in space to see its different faces. Applicability of data warehousing is constrained to some business scenarios. Data analysis has become crucial in variety of applications like non-business domains such as government, science, education, hospital, research, medicine etc. As the data is increasing, tweet data is stored in XML form and in XML ware house data are stored in facts and dimensions .Xml is flexible where ambiguities and inconsistent data are removed from the ware house and any time newly data can be introduced in to data warehouse. This cube view can do the aggregation of data.

Xiong Lui [6]has explained in his paper a text cube approach on social media analysis especially sentiment analysis. In data warehouse, data are stored in a multiple hierarchies and many dimensions and the queries are always accessing the data for better visualization. Data cube allows the data to be aggregated and viewed from multiple perspectives. The measures called as facts are the numerical values where we can do the numerical operations on the numeric data. Data cube has elaborated the structured data by representing it in a text cube format. These data in the table for measure has been collected from the original database. This text cube approach do text analysis capability for extracting HSBC measures from unstructured tweet stream. After viewing and analyzing tweets data, using cubes and charts data is also presented using the heat map. Each zone has been colored as blue and shows the degree of opacity which is directly proportional to the value of the measure. Now a days the research are concerned on how sentiments are expressed in online reviews and news articles. This is a social and cultural way for human to express their emotions. Emotions reflect how an individual sees the world and also how he reacts to the world in a particular event. Year's back cognitive psychologists didn't take sentiments of human in considerations. Understanding individual or analyzing the behavior of a group of people helps to predict the future and also to get the decisions for any business process. Emotion can be assumed not only non-verbally but also using text-based communication. In recent studies it has been found that human convey their emotions using text, blogs, emails, short messages and different textual conversation. If a person is already sad before any start of conversation,

the other participant can understand his negative emotion in text based communication. Emotions can be classified as negative positive or neutral. Suppose you have used some particular product and you wrote some reviews showing your emotions about that product brand in social websites. Company fetches all those reviews of the customers about its product in some particular area and do proper sentiment analysis which helps to take proper business decisions. Sentiments of each tweet is represented how much accurate the positive and negative sentiment it is and also the sentiment type is also shown. This data is put back into the database based along with the other partitioned labelled data. Dimensions and facts tables are now created based on twitter data. I have taken telecom industry data. Mondrian is an open source Online Analytical Processing server which is in java , which responds to the query fast enough to allow an interactive exploration of data , even there are millions of records, occupying gigabytes of data. Tomcat is an application server from the Apache software foundation that implements Java servlets and Web pages that include Java server Page coding. Tomcat helps to show the results of the open collaboration of the developers. Mondrian use OLAP4j API which is an extension of JDBC for OLAP applications that use XML provides for analysis. The cube schema is written in xml file where we mention the dimensions and the level of hierarchy of the dimensions we want to keep .and the measures along with the operations like sum, count etc. we want to display. This helps to create an OLAP cube. When we start Tomcat application, the Mondrian automatically generates MDX query that is multidimensional expression query. This MDX query helps to plot and to display the data in dimensions and measure.

Svetlana Mansmann [7]has explained in his paper how to create an OLAP dimensions in Semi Structured data. This work is related to integrating data warehousing and mining, OLAP for complex data and the social network data analysis. Data warehousing and Online Analytical Processing useful in different components of systems which helps in decisions support and business intelligence. It has been originated in the 90s where they have granted the permission for giving the access to the data to decision makers. Components that consist of for building these type of systems are databases and applications that offer the tools analysts which needs to support the decision making in the organizations. In an organization data is the more important assets, these assets can be stored in two ways one is operational system of records and another is data warehouse. Operational system records store the data about taking orders, signing up new customers and logging complaints. It can query one row at a time which is quite time consuming and not Efficient. On the other hand Data warehouse store the data so that it can get the data from it and do operations. It basically deals with the count of new

orders, why new customer signed, and why they have logged a complaint etc. It can access hundreds and thousands of row at a time. Data ware house is a specialization of database technology for integrating, accumulating, analyzing and also visualizing data from changed sources. It employed the multidimensional data model, which represents the data in a cube model which contains measures of interest. Data warehouse contains data that characterizes the business history of any company or organization. This ancient data is used for not only analysis; it also gives the provision of taking the business decisions at many levels something like strategic planning to routine evaluation of a discrete organizational unit. If the data in then data warehouse represents the data in a multidimensional database then data are stored in OLAP cube. OLAP tools give the benefits of creating online statistical reports by the means of query and the analysis of data warehouse information. Summaries are calculated using aggregate functions like AVG, SUM, MAX, MIN many more. There are such cube operations like slicing, dicing, drill-up, drill-down, roll-up, pivot etc. is used to explore multidimensional cubes. A structured model from the original tweet. The main motive of this step is to identify the available entities, their value domains, constraints and the relationships between entities. UML notation is used for the structured elements. The main class in this diagram is Tweet and User whereas rest of all the elements defines the relationship that relates either or both of them. User related information is the profile image, the location, the searches performed, notification received, users interaction. This interaction shows the statistics values as the counters on the followers and the following others, status updates, friends count. Tweet-related features comprise of location, the source of tweeting. The other users mentioned, media embedded and also the count of re-tweeting and favoring the tweet. Relationship between the users and the tweet can be authoring or retweeting the message or being mentioned in the message.

First twitter data has been fetched and it is kept in the efficient BaseX. Efficient base x in a database where huge amount of json file can be stored. Later this json file has be transformed into the xml file and stored in XML database BaseX. Then the data is represented in facts and dimensions and stored in Microsoft SQL server. From SQL server OLAP cube has been created by dragging and dropping the columns and rows. a structured model from the original tweet. The main motive of this step is to identify the available entities, their value domains, constraints and the relationships between entities. UML notation is used for the structured elements. Fig 3.1 shows the result of the relational mapping where a set of relations is linked by foreign key constraints. The main class in this diagram is Tweet and User whereas rest of all the elements defines the relationship that relates either or both of them. User related information is the

profile image, the location, the searches performed, notification received, users interaction. This interaction shows the statistics values as the counters on the followers and the following others, status updates, friends count. Tweet-related features comprise of location, the source of tweeting. The other users mentioned, media embedded and also the count of re-tweeting and favoring the tweet. Relationship between the users and the tweet can be authoring or retweeting the message or being mentioned in the message.

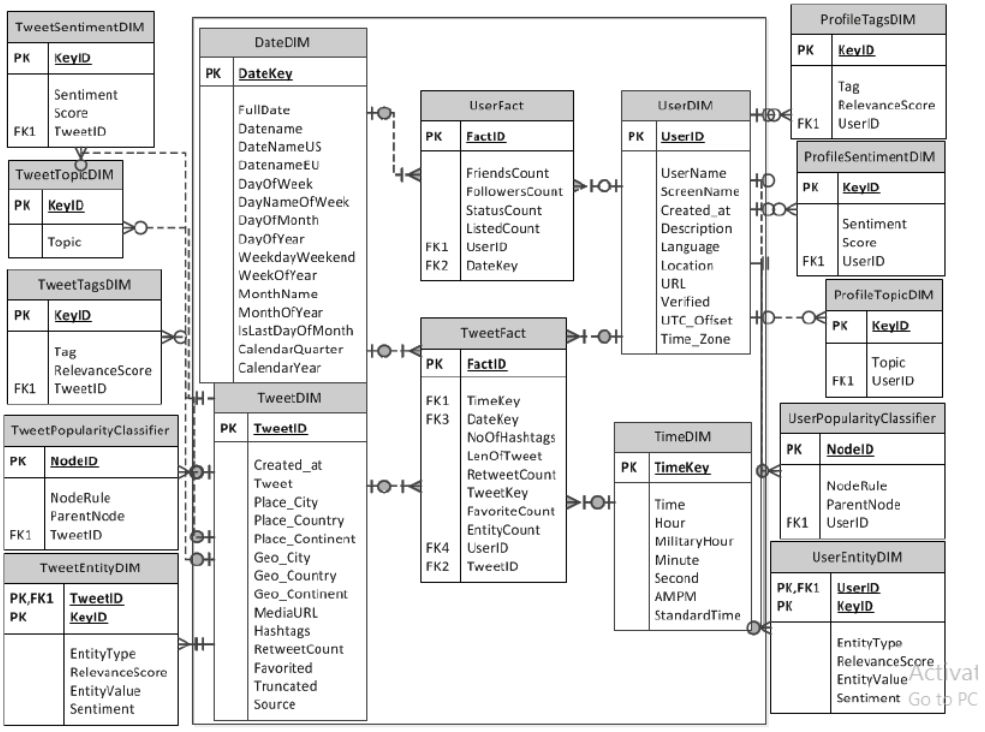


FIGURE 3.1: Relational View of Tweet Record

In this paper author[8] has gathered huge amount of dynamic textual data from the social media and using effective machine learning mechanism it has done sentiment analysis. It also has gathered the feedback on the attitudes and the opinion dynamic of the customers. Sentiment analysis methodology it has explained in Fig 3.2.

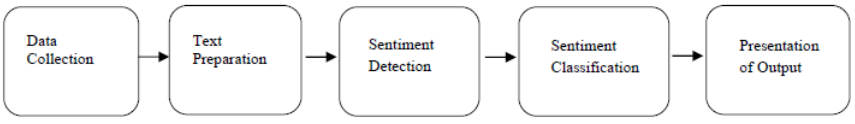


FIGURE 3.2: Sentiment Analysis Methodology

It has collected the data over the internet about people who has posted on blogs, forums, discussion boards and product reviews etc. As the collected data is huge, not organized, disintegrated and inconsistent.

After extraction of text make sure the text gives the information about the user, time, location and topic name. Then sentiment detection is done by extracting the data and using machine learning algorithm. Techniques are of different levels like unigrams, trigrams, lemmas and negation, then the sentiment is classified using positive, negative and neutral.

In this paper[9] due to the he amount of data all around the web sources like review, sites, blogs and new corpora, these are available in digital form where the new researchers are focusing on sentiment analysis. People are trying to invent a system which can classify and identify opinion and the sentiments as required. This type of sentimnet analysis can useful for predicting customer's preferences which is economic and valuable fro the research or the market analysis. Data sources from where data has been taken are blogs, review sites, dataset and the microblogging sites. Sentiment classification is done usong machine learning language that is suoervised learning where two documents are needed. One training set another is test ser data. Traing set is used for simply calssify and learn the characteristics of the documents in the test set data. There are many numbers of classifier Naïve Bayes, maximum entropy and support vector machine etc. Most wekk known machine learning languages are K-nearest neighborhood, ID3, C5, centroid classifier and N-gram.

In this paper author [10] has implemented sentiment analysis on three different techniques. HE has used rule-base classification, supervised learning and machine learning techniques, combined them and created a new method. He has used reviews of movie, MySpace social media data and reviews of the products as an input. The results shows the classification of micro and macro averaged F1. Micro averaging has a set of tables were the values are given two by two matrix.Each cell in that table gives a new column where it stores the sum of the document from that table. This new table gives the average value of new classifier. In macro matrix from a given table, new set of values are generated and each value gives the automatic recall for each categories of values.

We describe [11] Here author has described how to detect sentiment for short tweets and the unstructured texts.In tweets there are many types of Lexicon. These simply generates from the tweets with the hash tags, emoticons etc. To understand the tweets

and analyze the sentiments, this emoticons should be omitted from the tweets and get the structured tweets.

Author where[12] has considered newspaper and blogs where people express their feeling about any topic. this messages have been taken an scores are assigned and labeled whether it is positive, negative or neutral.They have used machine learning language where they have used corpus for training set data. Corpus is quite large for news and blogs. Semantic orientation of words are analyzed where word 'and' have some polarity where 'but'having opposite polarity. little lists of words have been created for much more accuracy of sentiment. Author has created different lexicons for various dimensions.Like for general, health,crime,sports,business,politics and media.

3.2 Critical Review of Literature

Author Killian Thiel[13] has put an approach towards creating usable customer intelligence from social media data where he has explained about the social media score board and predictive analysis techniques. Social media score boards are the tool for cloud based applications where data is collected from different sources and it constantly captures the new data as it occurs. It provides the user to look across the wide range of channel choices. This display is done using visual scorecards where it combines different graphical and tabular techniques for delaying the summarized information.For further details, drill down can be down to look right through the data.

Another is predictive analysis technique starts creating new facts in the social media date. API allows to fetch the dynamic data and these data uses for predictive analysis. Predictive analysis are the sentiment analysis or the network analysis. He has also explained a huge number of data mining tools which are open source and commercial both. This type of tool is generally used in the web age and on the web applications which serves to get the data from the cloud based application. Channel reporting tolls is one kin of tool which particularly used for getting the immediate on the focused activity and the activity which we are looking for some changes either in real time or in a fixed time. Among the one of the tool is social media scoreboards. This score boards is cloud based applications which captures all kind of data sources from the social media and stores the data which includes communities, and blogs. This data actually clears the criteria and enter into the specific channel for more analysis work. Predictive analysis is also an analysis technique where we use trend view. Trend view is where we predict

the sales or the data depending on the previous data. This huge amount of data can be used and utilized in the organizations well fare. Thus companies are trying to monitor and understand, analyze and measure the eventually growing space of the social media. It explains importance of social media business for its measurement and the limitations of the web analytics and its solution, capturing social media content and using it in individual business context. Web analytics is the process that measure, collects and analyze and also reports the internet data for the whole purpose of understanding and using the web. It helps whole measuring traffic and also provides us an effective tool for business and market research. Here author has taken the data sources as the social media like, earned social media, bought social media, own social media and the other media from where he has taken the data sources.

Later he has put the data as an input in the property inventory or using the mining keywords. Many natural language processing tool kit have been used for analytics and processing of data like NLP algorithm, text miner, crawlers and text extraction. Through this data is processed and quantitative and qualities report is made after that. Then business decision makers actually do the quantities insightful job.

Author Weiguo Fan [14]et.al has explained the power of social media analytics in his paper. He has explained social media analytics into some process. First is capturing the data. Gathering the data from various sources, preprocess the data, that is removing the noisy data, and extract the information from the data. Next is performing analytics that is opinion mining, sentiment analysis, opinion mining, topic modelling, social network analysis, trend analysis etc.

Then it comes presenting the data. Presenting refers to summarizing the data and evaluating the findings from the stage 2 and also present the filings. First it captures the data which is related to the business logic. It identifies hundreds and thousands of information lying on the social media platforms relates to their activities and interests. API, web crawling allow to get the dynamic data directly from the websites.

Next is understanding the data.once business related data has been captured, this noisy data and unstructured data is analyzed. Rule based text classifiers ad more sophisticated classifiers trained on the labelled data and may be used for the cleaning of data. This cleaning involves many related statistical methods and their techniques. This stage provides the understanding between the users sentiment, how they are feeling about any company, the brand, the product they have used etc. many valuable trends, metrics, purchasing history and also the prediction can be done.

The last stage is the presentation area where different analytical results from different analytics are summarized, explained, calculated, and shown to the consumers to make the easy for understanding. This type of techniques are useful to present the information in front of them. There are many sophisticated visual representation of data that present the data in the form of different pattern.

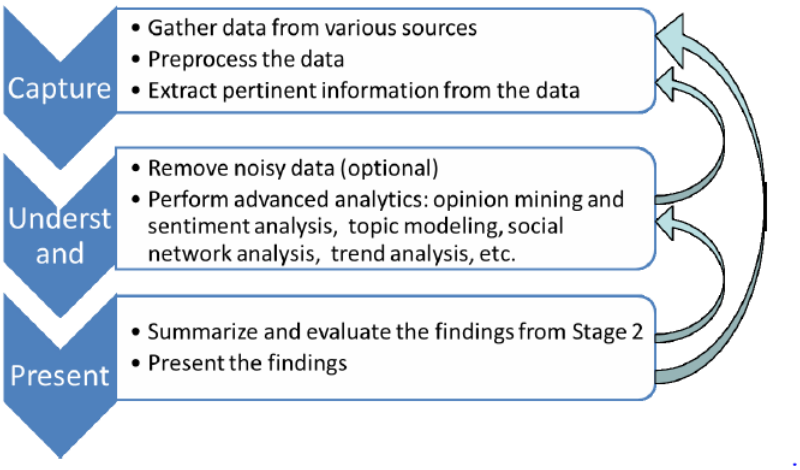


FIGURE 3.3: Steps for Data Analytics

Bogdan Batrinca [15] has explained many tools and techniques to do social media analytics work. This research requirements of the papers are data, analytics and the facilities. Data can be of many type and can be obtained from everywhere, like social network media, news data, public data and programmable interfaces. Another is the analytics where researchers tempt to do analytics using a programmatic way like mat lab, java or python. Non programmatic way of doing analytics is analytics dashboards. This requires a deep access to the raw data. Another is the holistic data analysis where tolls are used to combine huge number of data from the social media and their datasets. Data visualization is also important where researchers use tool which helps to represent the data in a graphical form and also it helps to understand clearly.

Facilities are huge number of data is stored that has been collected from twitter. The data is stored and it has been very useful to the researchers for their further project. Challenges he has taken for implementing this paper, first is scraping of data, social media data is copied using may web scrapping tool. Some microblogging web sites give the chance to get the dynamic data, data though the API, it creates OAuth that verifies and gives the permission to the sure to extract the information. After that data cleaning- this stage understands the unstructured data that is of high frequency. It uses

many tool for understanding the data. Once business related data has been captured, this noisy data and unstructured data is analyzed. Rule based text classifiers and more refined classifiers trained on the labelled data and may be used for the cleaning of data. This cleaning involves many related statistical methods and their techniques. This stage provides the understanding between the users sentiment, how they are feeling about any company, the brand, the product they have used etc. many useful trends, metrics, purchasing history and also the prediction can be done.

Data analytics, where sophisticated analysis work has been done like sentiment analysis, opinion mining. Challenges comes due to the foreign languages, foreign words, slang, spelling error etc. many social media uses analytical dashboards ,which doesn't need and program and it's good to present data where deep access to the raw data is needed. For better visualization, data is presented in a graphical form.

Nafees Ur Rehman et.al[16] has explained how powerful the OLAP when it comes to solving a specific Twitter-related analysis tasks. Author has extracted the twitter content for three hours on pertaining to the earthquake in Indonesia on April 11, 2012. The task actually related the understand the twitter in a better way while there is an emergency. Famous microblogging sites such as twitter has established and it has changed the way of users interact with each other and share digital content. It has grown more than 465 million accounts. While earthquake total number of tweet count has be calculated across the world plus the no. of tweet count and retweet count has been calculated in during the Timeline of the earthquake in Indonesia. All the numbers and the figures are represented by Analysis Services Toolkit of the Microsoft SQL server. He has explained that it offers a user-friendly interface and interactive visual exploration of data. The analyst move through the data which is characterized in facts and dimensions to views the cubes properly and drag the elements. Any other kind of data analysis job can be performed in the similar fashion. Data in the data warehouse is well structured and organized. An entry in fact table consists of measurement, metrics of facts or events.

Its location is the center of the star schema or the slow flake schema which is surrounded by the dimension tables. The fact consists of some measures like performance indicators with its whole description in the dimension table. The representation of facts table and its associated dimension and the classification hierarchies is called multidimensional data cube. In twitter star schema dimensions can be user details, date, location, product details, zone, and sentiments. Sentiments are the linguistic feature that can be added in the dimensions. Based on the Star Schema, data cube architecture can be designed to

let the users to calculate the totaled statistics of sentiment related measures along with the different dimensions like time and locations etc.

Efthymios Kouloumpis et.al[17] has referred one paper name Twitter Sentiment Analysis: The Good the Bad and the OMG! Where he has explained about the linguistic analysis that is sentiment analysis. Now a days the research are concerned on how sentiments are expressed in online reviews and news articles. This is a social and cultural way for human to express their emotions. Emotions reflect how an individual sees the world and also how he reacts to the world in a particular event. Year's back cognitive psychologists didn't take sentiments of human in considerations. Understanding individual or analyzing the behavior of a group of people helps to predict the future and also to get the decisions for any business process. Emotion can be assumed not only non-verbally but also using text-based communication. In recent studies it has been found that human convey their emotions using text, blogs, emails, short messages and different textual conversation. If a person is already sad before any start of conversation, the other participant can understand his negative emotion in text based communication. Emotions can be classified as negative positive or neutral. Suppose you have used some particular product and you wrote some reviews showing your emotions about that product brand in social websites. Company fetches all those reviews of the customers about its product in some particular area and do proper sentiment analysis which helps to take proper business decisions. The tweet when it is cleaned is fetched to do sentiment analysis. Sentiment analysis is done using natural language tool kit (NLTK) or SVM (Support Vector Machine) etc. which all is predefined algorithms. NLTK has its own corpus name NLTK data which has predefined training set and also while using NLTK we need to have training set and test set. Training set trains the test set data. Test records learn from the training set data and then predict. For sentiment analysis three training data set should be there one negative, one positive and one neutral which will classify the data in the test set after learning from the training set. Each review will be labeled as positive, negative and neutral.

Chapter 4

SOCIAL MEDIA ANALYTICS

4.1 Methodology for the Study

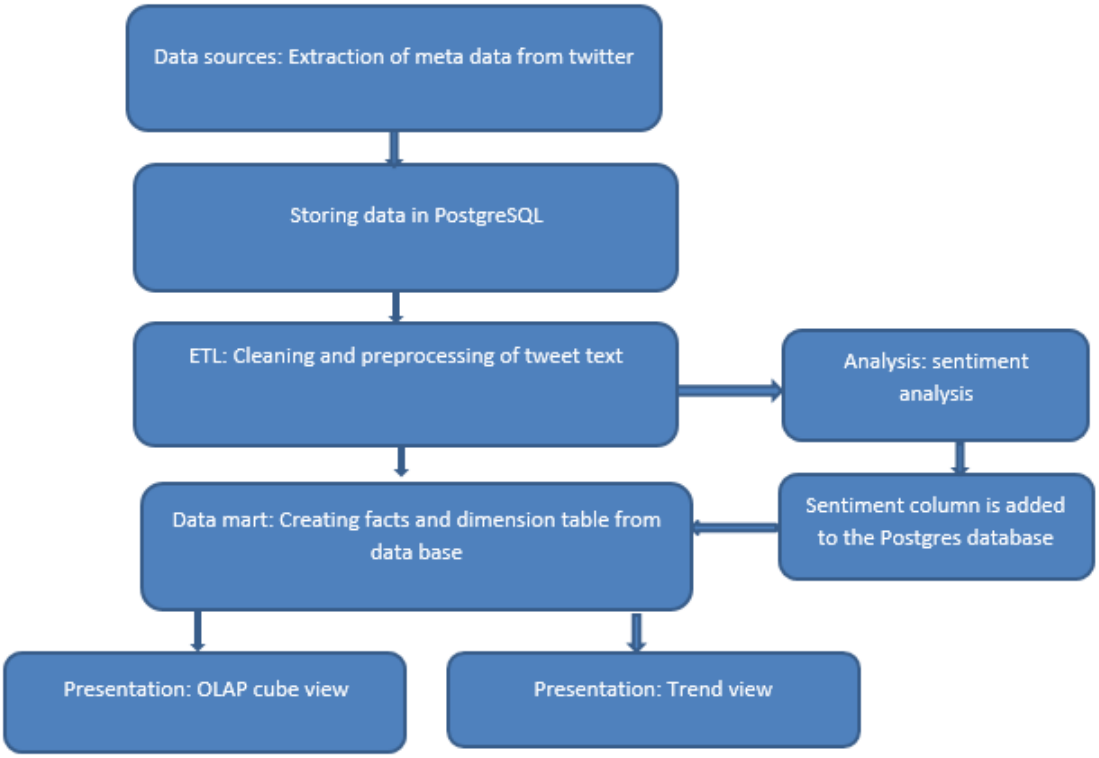


FIGURE 4.1: Work Flow Diagram

4.1.1. Data sources

Tweet data is unstructured data twitter allows to post tweets of 140 characters. In 140 characters if becomes tough to explain the actual feelings. People use, icons, hashtags etc. to express their feelings. This unstructured data is in json format. In twitter there is a particular form of data that is, data is presented in the form of Meta data. Data, inside another data. Twitter allows users to extract dynamic data direct from the twitter. Daily 500 million data twitter, generates, and users can download the data using twitter API. API is application program interface, it is a set of routines, tolls, rules and protocols for building any software application. This explains how to interface with the programming graphical user interface that is (GUI) and how the software components should react. API creates a program by making each and every block separately and later program joins the every block. API can be of many types of API depends on the operating system, applications and websites, windows. When someone copy and paste the code into some other application, this API only helps to work with that code. Twitter supports many API. The public twitter API consists of a REST API and another is the streaming API. Many application has been built by missing and matching many API by the developers. The streaming API produces low latency high-volume access to tweets. REST API gives the opportunity programmatically to read and write the twitter data. . Author of any new tweet read, details of the author's profile, followers count etc. Can be known. This API understands the applications and the OAuth that gives in response twitter json data. In Fig.4.2 i have explained API which is Streaming API. Streaming API frequently delivers responses to the REST API as it wants to have the HTTP connection long lived. It receives the latest new updates of any tweet, matching the query and staying in sync with the user profile and more.

First Twitter Application is created where we create an OAuth. OAuth is to have secure and authorized end point connections. In twitter website, when creating an twitter API, we fill a form where we add website name, and the name of the API then OAuth is verified by creating four keys. These keys are consumer keys, consumer secret key, and access token key access token secret key. While extracting the data in a programmatic way from twitter, these are verified by the twitter, where the OAuth actually works. Then Twitter Meta data will automatically downloaded. We can download tweets based on the topic we want and the no. of days, from which date data, or the particular date data

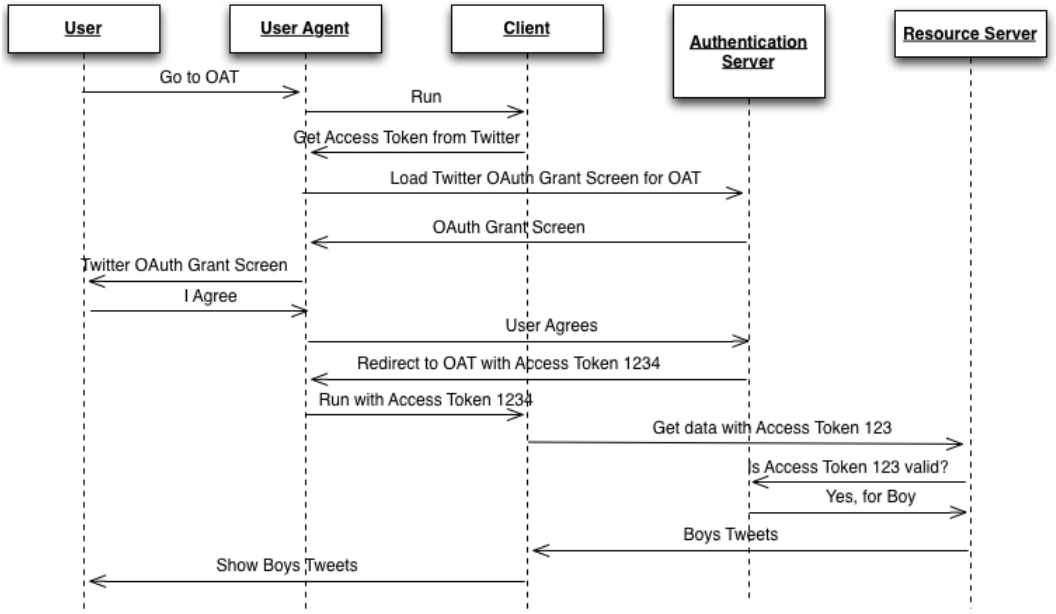


FIGURE 4.2: How twitter API works

4.1.2. Storing the data

After extraction of data from twitter, the data is stored in PostgreSQL. Postgresql is called as postgres, is an object relational database management system. As it is a database its basic work is to store the information, securely, supporting huge amount of data, quickly responding the query while manipulating or fetching and any updating of the data in the database and also allows to access the data by other software applications. It allows may parallel operations from single user, single machine accessing to a number of users, accessing with the internet connected applications. PostgreSQL is better than other databases as it has a better support than the other proprietary vendors, it has the significant saving on staffing cost like, it needs lower maintenance and less requirements but it gives much secure and stable good performance . This database has the reliability and the stability that easily data won't get lost or crashed after several years of hard and more active operations. PostgreSQL allows source code without any charge or in minimum effort. Postgres supports most of the operating system like, UNIX, windows etc. It is compatible with many systems without any effort. It also supports multiple rows and stores huge amount of database. Postgres supports high quality of GUI tools which helps in commercial applications.

While programmatically fetching the data from twitter, it automatically stores in the postgreSQL. As I have fetched twitter Meta data specifically, instead of json data, so the data is stored in a proper structured column wise. Before that, I have created a

table of the columns I want in the postgreSQL database, and in the same order I have mentioned so that data automatically inserts into column without overlapping with each other.

The data stored in PostgreSQL database in the same order as follows:

1) sid(integer)
2) tweet text(text)
3) tweet fr text(text)
4) created date(date)
5) screen name(text)
6) is retweeted(boolean)
7) reply to sn(integer)
8) profile location(text)
9) favorited(boolean)
10) favorite count(integer)
11) reply to sid(integer)
12) id(integer)
13) reply to uid(integer)
14) retweet count(integer)
15) retweeted(boolean)
16) status src(character varying)
17) network provider(text)
18) followers count(integer)
19) following count(integer)
20) created time (time with time zone)
21) sentiment(text)
22) score(integer)
23) pos(text)
24) topic name(text)
25) len of tweet(integer)
26) tweet count(integer)

4.1.3. ETL

ETL process is one of the important task in Data staging area. Extraction, load and transformation is the stage. This is the stage where data gets preprocessed. ;data collected from the twitter, contains hash tags, symbols, icons. This needs to remove for the

analysis of data. Suppose you are at a restaurant. People who comes for the restaurant is not supposed to see what is going on in the kitchen room. Same as here, users or the decision makers suppose not to see what is going on is the ETL stages. As raw daya has been inserted and the preprocessed ready data will be the outcome for this stage. Extraction is the first step where data is extracted from the postgres. Tweet text is extracted from the " tweet for text" column. The tweet contains misspellings, errors, domain conflicts, these are removed through python program and also symbols are removed. This stage do lot of transformations. Transformations like applying business rules,cleaning of data, filtering the data on some particular keyword,splitting of data,joining the data together,applying any kind of simple or complex data validation.Fig4.3 is where the final step is loading the data. The data is loaded in the excel sheet for further.

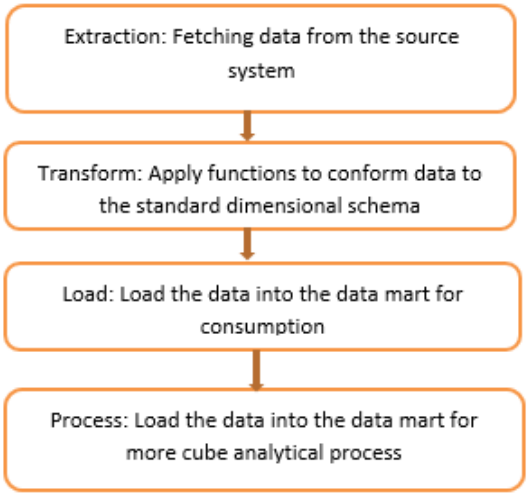


FIGURE 4.3: ETL process

4.1.4. Analysis

In this gradually growing world of social media, analysis has become a major part. User's post has been collected by the organizations and they have taken an approach to listen the texts and analyze them for their organization's betterment. Customer has become the most popular sets in the defined market. First step of analysis is listening to the data. It tried to identify and collect those data particularly on which analysis can be done.

I have taken network provider data where I have mentioned while fetching the tweets that from which date to what date i want the tweets to collect and on what topic it should come. Analyzing the collected data and to understand is the customer is satisfied or, if he is facing any kind of problem. Sentiment analysis is one kind of linguistic analysis. This is a social and cultural way for human to express their emotions. Emotions reflect how an individual sees the world and also how he reacts to the world in a particular event. Year's back cognitive psychologists didn't take sentiments of human in considerations. Understanding individual or analyzing the behavior of a group of people helps to predict the future and also to get the decisions for any business process. Emotion can be assumed not only non-verbally but also using text-based communication. In recent studies it has been found that human convey their emotions using text, blogs, emails, short messages and different textual conversation. If a person is already sad before any start of conversation, the other participant can understand his negative emotion in text based communication. Emotions can be classified as negative positive or neutral.

Suppose you have used some particular product and you wrote some reviews showing your emotions about that product brand in social websites. Company fetches all those reviews of the customers about its product in some particular area and do proper sentiment analysis which helps to take proper business decisions. For sentiment analysis I have used Naïve Bayes classifier. It is a natural language tool kit.In Fig 4.4 Naïve Bayes classifier is a simple probabilistic model which is based on the Bayes rule with the strong independence assumption. This Naïve Bayes model gives a conditional independence assumption which has the class positive and negative bag of words which are conditionally independent to each other. This type of assumption doesn't harm the accuracy of the text classification but speed up the classification. If we use simplifying conditional independence assumption then the given class that is positive or negative words will conditionally independent of each other.

After sentiment analysis the data is posted back to the postgres database from the excel sheet. We have user sql query function one inserting data in the postgres. We alter the table by adding one extra column in the postgres main database. And added the data from the excel sheet.

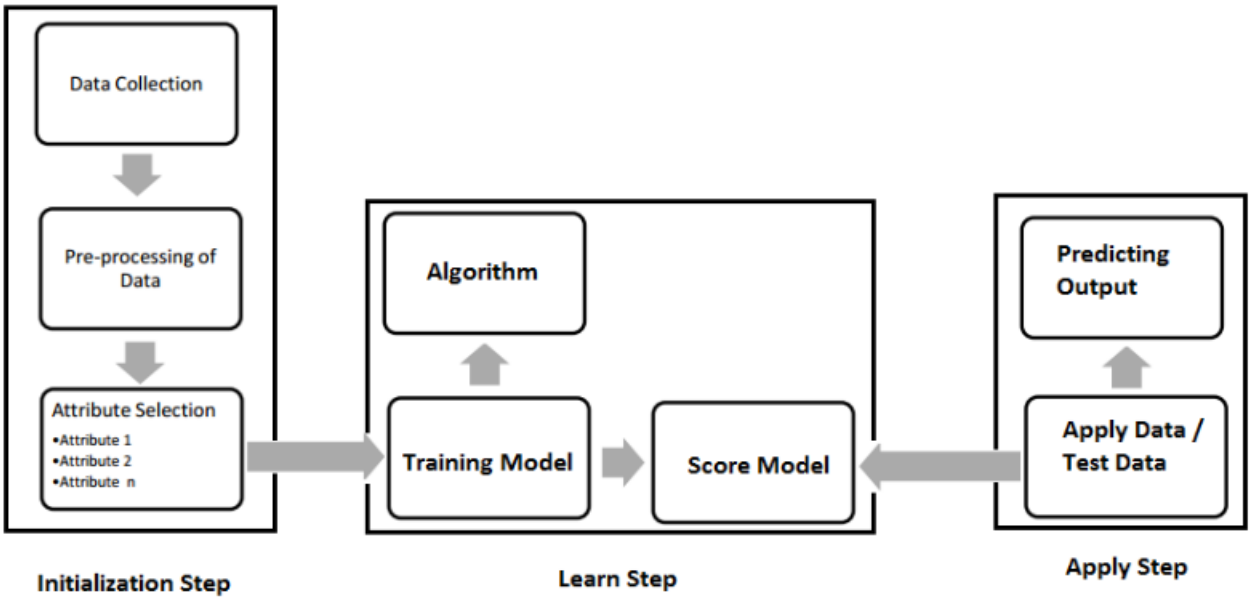


FIGURE 4.4: Using Naive Bayes Algo sentiment analysis

4.1.5. Data Mart

Data mart is a small portion of the data warehouse. As data warehouse holds big amount of data but data mart stores the subset of the data based on a single topic. Data mart can be many from a single data warehouse. It's based on the requirement of the organization. Data mart is created using facts and dimension table, it is known as conformed.

A fact table is the table which is the basic requirement for creating the dimensional table. As fact table stores all the numeric value for the measurements used for the business process. This fact table reduce the duplicity of data across the table in the data mart. Many fact tables can be created from a list of dimension table. One data mart can have many number of fact tables. Fact table does the calculations. Mostly it is used for the business purpose. Where the query hits the fact table in the database, it retrieves not only one single row, hundreds and thousands of rows are retrieved. Another important thing is it adds up the data and show the agg value. The fact table can't be represented as a zero value because, if we consider zero value, it will be spreader in the fact table everywhere. Fact table satisfies referential integrity when all the primary keys of the dimension table matches with the foreign key of the fact table. The fact table is being

accessed via the dimension table. Every fact table has their own primary key, and also many foreign key. That key is called composite key. Fact table represents many to many relationships. Fact table guarantee uniqueness between the rows.

Dimension table is the main part of the fact table. It contains the attributes or the description of the business process. This table have more number of columns and each dimension table has a primary key which helps in grouping or reporting labels. Suppose someone wants to see the following count for the particular date, day, week day and the number of year of that day. This can be possible in dimensional table. Source table is named as cubeview data columns names as follows,

1) sid(integer)
2) tweet text(text)
3) tweet fr text(text)
4) created date(date)
5) screen name(text)
6) is retweeted(boolean)
7) reply to sn(integer)
8) profile location(text)
9) favorited(boolean)
10) favorite count(integer)
11) reply to sid(integer)
12) id(integer)
13) reply to uid(integer)
14) retweet count(integer)
15) retweeted(boolean)
16) status src(character varying)
17) network provider(text)
18) followers count(integer)
19) following count(integer)
20) created time (time with time zone)
21) sentiment(text)
22) score(integer)
23) pos(text)
24) topic name(text)
25) len of tweet(integer)
26) tweet count(integer)

Score is an attribute which is calculated based on ((retweet count*80)+(favorite count*20))

Length of the tweet is calculated in the sql. And then this value is added back to the postgres with thee source table.

Fact table is surrounded by the dimension tables. In my project I have taken 6 dimension tables. Date dimension table represented a date dim as follows,

Attributes are:

1) date key(serial PRIMARY KEY)
2) date (integer)
3) year(integer)
4) month (integer)
5) month name (text)
6) day (text)
7) day year (integer)
8) week day of name (text)
9) calendar week (text)
10) formatted date (date)
11) quartal (varchar)
12) year quartal (varchar)
13) year month (varchar)
14) year calendar week (varchar)
15) weekend(text)

As we have extracted created date. From there through sql query in postgres we have extracted and put these attributes columns for better business analysis.

Network provider name dimension table named as network provider. Attributes are:

1) np id (serial PRIMARY KEY)
2) network provider(text)

Time dimension table named as time dim Attributes are:

1) time key (serial PRIMARY KEY)
2) created time (time with time zone)
3) minutes (integer)

4) seconds (integer)
5) miliseconds (integer)
6) hour (integer)

Same as created date, created time attributes was extracted and these columns are put using sql query.

Sentiment dimension table has been created named as sentiment dim Attributes are:

1) sentiment id (serial PRIMARY KEY)
2) sentiment (text)

User dimension table named as user dim Attributes are:

1) user id (serial PRIMARY KEY)
2) screen name (text)
3) lang (text)

Tweet dimension table named as tweet dim Attributes are:

1) tweet id (serial PRIMARY KEY)
2) created date (date)
3) tweet text (text)
4) retweet count (integer)
5) favorited (boolean)
6) favorite count (integer)
7) status src (varchar)

For Log in two tables have been created,table users

1) user id
2) user name
3) passwords
4) superusers
5) rights

Another table is user login details table,

1) id
2) ip
3) user name
4) log date time
5) application details
6) log out time
7) duration

Fact tables are for cube view Tweet fact table references as tweet fact Columns are:

1) fact id (serial PRIMARY KEY)
2) date key (FOREIGN KEY)
3) sentiment id (FOREIGN KEY)
4) time key(FOREIGN KEY)
5) tweet id (FOREIGN KEY)
6) len of tweet
7) score
8) tweet count

Another table is user fact table referred as user fact Columns are:

1) user fact id (PRIMARY KEY)
2) user id (FOREIGN KEY)
3) date key (FOREIGN KEY)
4) sentiment key (FOREIGN KEY)
5) time key (FOREIGN KEY)
6) np id (FOREIGN KEY)
7) following count
8) followers count
9) tweet count

Fact table for trend view Columns are:

1) agg fact id (PRIMARY KEY)
2) date key (FOREIGN KEY)
3) sentiment key (FOREIGN KEY)

4) time key (FOREIGN KEY)
5) np id (FOREIGN KEY)
6) user id (FOREIGN KEY)
7) following count
8) followers count
9) retweet count
10) year
11) month
12) network provider
13) sentiment
14) screen name

4.1.6. Presentation

Representation is this facts and dimensional table is the cube view and the trend view.

OLAP cube view: An OLAP cube is a multidimensional database that is basically used for the data warehouse and the online analytical process applications. OLAP cube is the way where to store the data in a multidimensional form for mostly reporting or taking business decision purpose. The data that represents the OLAP cube is categorized by the dimensions. For creating and accessing the OLAP cubes, the code is written in XML and the query is multidimensional expression (MDX). This language has been adopted by many vendors such as Mondrian. OLAP cube can be made using traditional database but the dimensions must be proper. The OLAP cube is basically used for report making in the business logic where thousands and many rows can be displayed and show. It does the analytical work. Every cube is made of schema. This schema file is a joined table in the data warehouse from which you gets the actual data from the source table. Cube is created using the fact and the dimension table. If any company wants to see the aggregate value of all the following count of that specific network provider, and compare it with the other network provider, then cube view is the best representation of data where it can be shown properly. In my project I have created two cubes: One is using tweet fact table and another is using user fact table. Using tweet fact table you can plot a cube. The measures and the dimension you want to add should be properly written in xml cube schema. Suppose, in a particular date, how much the sum score of all the positive tweets is. For some particular date, how many much is the score of positive, negative and the neutral tweets. Like this many ways cube view data is view and analyzed by the decision makers.

Trend view: Trend view is one type of analysis where it allows the business users to predict the future about what will happen by seeing the previous value or the history. Trend analysis mainly deals with the historical data. It forecast based on the time span like 3 months, 6 months or 12 months. Business people can be the changes in the market by seeing the trend. This a type of technical analysis which predicts the future movement of the important measures of the organization. It gives the idea about any organization and also helps the organization to take actual correct steps at a correct time. This future prediction helps to understand and guess the future based on the uncertain events in the past. In my project I have created a table name agg fact table from there I have made this trend view. The measures are following count, followers count. On a particular time, of a particular network provider, what is the following count for 3 months will be shown in bar graph. You can see the color and the bars and also the color when there is a change in the data of the next month's respectively. Tomcat is an application server which hosts the webpages. Tomcat is installed in the computer which will hosts the user interface. Mondrian is an OLAP server. It is written in java. This executes the query very fast which is created by the user and then it creates MDX query on their own. Cube schema is written in XML file and saved as .xml. Proper dimension and fact table mapping should be done to view the data in tomcat. Multidimensional query plots the cube taken data and the dimension and measures name from the cube schema. With the help of dimension and fact table, same way in tomcat trend view is shown and analysis is done.

4.2 Analytical Work

4.2.1. Sentiment Analysis

In my project I have collected the raw tweets saved in the postgres. This data I fetched and cleaned using in a programmatic way. This cleaned data is further processed for the sentiment analysis. I have used python code for Naïve Bayes algorithm to do the sentiment analysis.

Two training sets I have created. One is positive words and another is negative words. These two files will train my text set data. This files have the collection of words which has the probability of occurring in the tweet text. In the tweet text each word is compared with the words in all the class files that is positive and negative. This do the comparison of the single word against all class file. According to this method, this

word is compared and the accuracy of positive or the negative value is found for the whole text. This accuracy can be more accurate . all depends on the class file which you have taken. To increase the accuracy of the tweet text your, class files should be more proper and specific. Those training sets actually trains the data, and by that test set data learn and the accuracy thus calculated. For better understanding and analysis, the sentiment accuracy of positive , negative is labelled as positive tweet , negative tweet or the neutral tweet. Before that negative, positive values are loaded into the source table. This source table where extra column has been added and this data is imported in the mention column. Using query we mention if the particular tweet is positive, negative or neutral. If the accuracy of negativity of a tweet is between 1 or 2 then the tweet is negative. Same as if the accuracy of the positivity of a tweet is between 1 or 2 then the tweet is positive else neutral. Like this way the sentiment is named and for each row the tweet is labelled as positive, negative and neutral tweet. This helps to understand the sentiment of the tweet like if it is positive or negative or neutral. It also helps and good to represent while doing analytical job. We can understand for a particular network provider, negative tweet people has how many following count and followers count. On a particular, time or date in neutral count or positive count tweets how much is the actual score of the tweet etc.

4.2.2. Cube View

From the main source table we have created 6 dimension table. Time dim, date dim, network provider, sentiment dim, user dim, tweet dim. And two fact tables tweet fact and user fact. This fact tables has the numeric values and it has the foreign key which is the primary key of eeach dimension table. This fact table is created by the query in the postgres. This query is inner join query which joins the dimension tables we want mapped with the main source table and then this mapping is created and inserted into the fact table. For mapping the data, the number of rows increases because for these many tables, for each combination of data, new rows has been created. Cube schema is created in xml. This is an XML file which has been created helps Mondrian to generate the MDX query. This MDX query creates the cube like which all the dimension and measures should be visible. In the cube schema we have mentioned which is the fact table needs to access for plotting a particular cube view. Suppose I have taken tweet fact as my fact table. This tweet fact plots on a specific date, of that specific month, of positive sentiments data how much the score is. If we consider user fact we can analyze for a particular network provider, for some time period how many followers count and

following count is considered. These all the analytics task has been done using this cube view data.

4.2.3. Trend View

The many task of the trend view is first understands the data, the content like what all will be the dimension and the fact table and how this table will be mapped in the cube schema to the trend view data. For minimum of 3 month gaps we can see the changes of the measures really visible. So, according to the collection of data, when we see the changes in the data between the time periods, this helps to analyze the future outcome. People can guess based on the uncertain history. It's all about data. For accurate analysis your historical data should be proper. In trend view the green bar shows if there is a change in the first or 2nd month, then the highest value will be marked as green. First month is market as green. If there is no change then all levels will be down. If there is at least 2 percent change I data, then the highest vale bar is shown as green else it will show as red if there is 2 percent decrease of value.

For trend view I have taken aggregate fact table. For analysis job, on a particular time, for a specific network provider, how the following count and the followers count are changing for last three months. If you are available with more historical data, then you can predict the data by seeing the data for last 1 year. There will be more accurate chance of prediction than 3 months.

4.3 Design

-

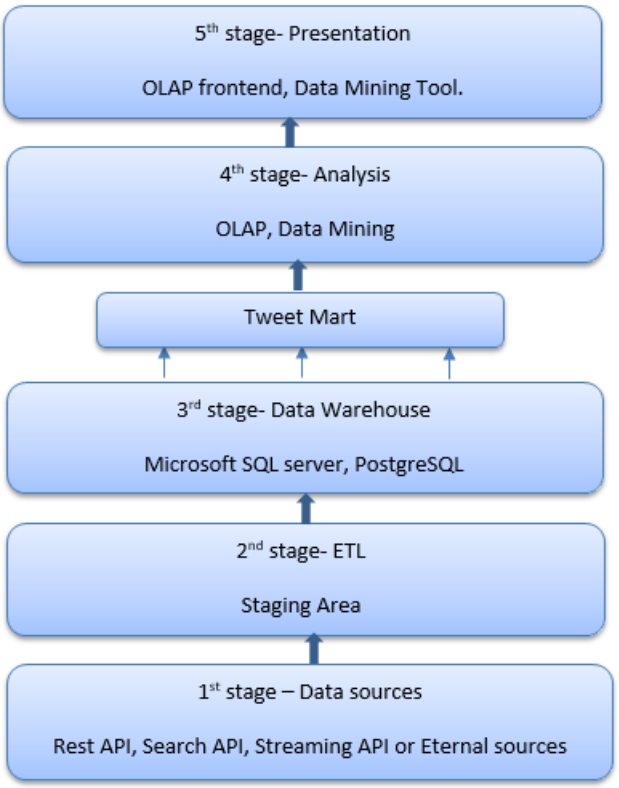


FIGURE 4.5: Twitter Data Warehouse

4.3.1. Twitter Data Warehouse

In Fig 4.5, it explains the architecture of the Twitter Data Warehouse.The very first step is the Data sources.that is the Twitter streamed data which is fetched using twitter API or REST API. After cleaning it goes through the last stage of ETL process, where data is loaded in the Data Warehouse. Sentiment is one of the most social and cultural dynamics for humans. It gives the sense of the emotions where we can get the positive, negative and the neutral sentiments of people based on their review. The data is represented using facts and dimensions which stores in a database as a cube and creates OLAP cube.In representation area, the data can be accessed using many tools and according to our need, tweet structured data is viewed in a cube format,and prediction analysis can forecast the future,using dimension and facts.

4.3.2. Dimension and fact tables

TABLE 4.1: Cube view data I

1	sid	tweet_text	tweet_fr_text	created_date	screen_name	is_retweeted
...700	...700	...700	...700	...700	...700	...700

TABLE 4.2: Cube view data II

1	reply_to_sn	profile_location	favorite_count	reply_to_uid	reply_to_sid	retweeted
...700	...700	...700	...700	...700	...700	...700

TABLE 4.3: Cube view data III

1	retweet_count	status_src	network_provider	followers_count	following_count
...700	...700	...700	...700	...700	...700

TABLE 4.4: Cube view data IV

1	len_of_tweet	pos	sentiment	topic_name	score	tweet_count
..700	..700	..700	..700	..700	..700	..700

TABLE 4.5: Time Dimension

	time_key	created_time	minutes	seconds	miliseconds	hour
1	1	17;58:16	58	16	16000000	17
2	2	03:22:24	22	24	24000000	3
3	... 615	...615	..615	..615	..615	..615

TABLE 4.6: Date Dimension I

1	date_key	date	year	month	monthname	day	dayofyear	weekdayname
..366	..366	..366	..366	..366	..366	..366	..366	..366

TABLE 4.7: Date Dimension II

1	calendarweek	formatteddate	quartal	yearquartal	yearmonth	yearcalendar
...366	..366	...366	...366	..366	...366	..366

TABLE 4.8: Network Provider Dimension

	np_id	network_provider
1	5	Maxis
...4	..4	..4

TABLE 4.9: Sentiment Dimension

	sentiment_id	sentiment
1	16	Positive
2	17	Negative
3	18	Neutral

TABLE 4.10: User Dimension

	user_id	screen_name	profile_location	lang
1	1	rcs963	Caen	en
...660	...660	..660	...660	..660

TABLE 4.11: Tweet Dimension

1	tweet_id	created_at	tweet_text	retweet_count	favorited	favorite_count
...333	...333	...333	...333	...333	333	...333

TABLE 4.12: Tweet Fact

1	fact_id	date_key	sentiment_id	time_key	len_of_tweet	tweet_id	score
..5387	..5387	..5387	..5387	..5387	..5387	..5387	..5387

TABLE 4.13: User Fact I

1	user_fact_id	user_id	date_key	sentiment_id	time_key	tweet_id	np_id
..5387	..5387	..5387	..5387	..5387	..5387	..5387	..5387

4.3.3. UML diagram of Twitter Data Warehouse

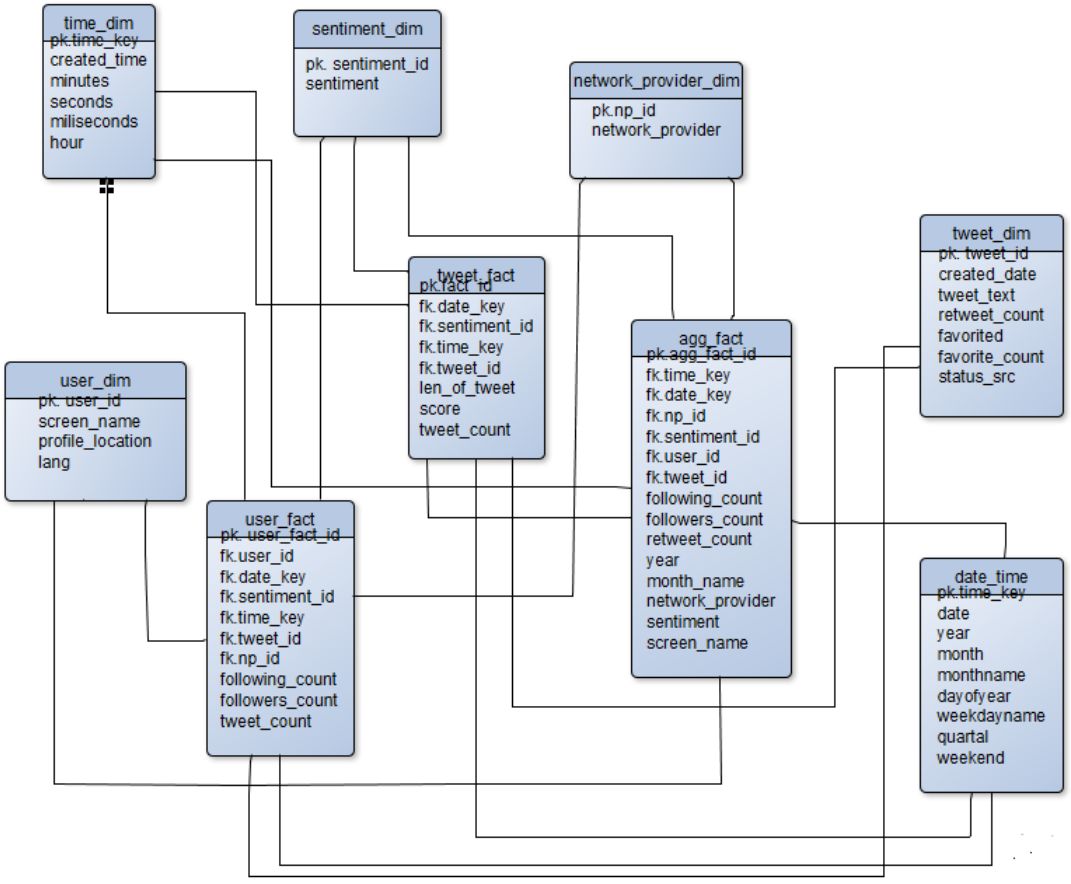


FIGURE 4.6: Star Schema model with facts and dimension table

Star Schema model of the Twitter Data warehouse where all the fact table are surrounded by the dimension table. The primary key of the dimension tables becomes foreign key in the fact table.

Chapter 5

RESULTS AND DISCUSSIONS

5.1 Results & Discussions

This project Social Media Analytics has carried out a number analytical task such as OLAP cube view and the trend analysis. Based on the data mart where data are in facts and dimensions, this representation and visualization of data has been implemented.

FIGURE 5.1: Front page of my application.

- Cube View:

Cube schema has been written in XML where the number of dimensions and measures and the operations have been explained properly.

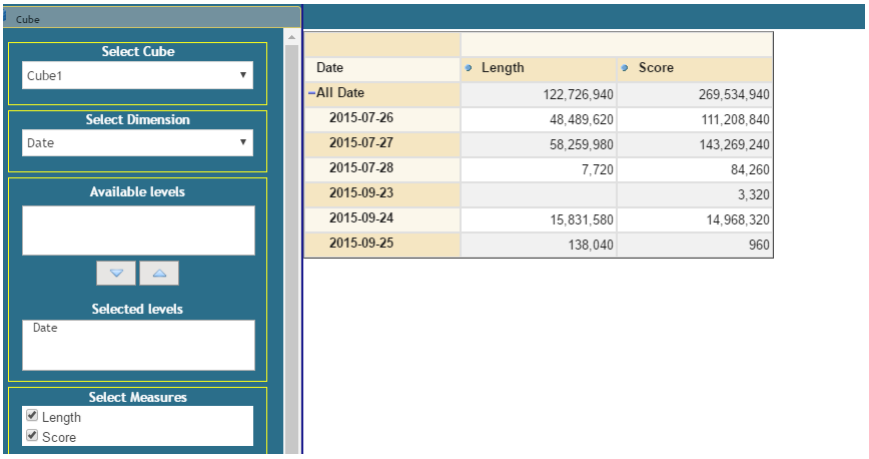


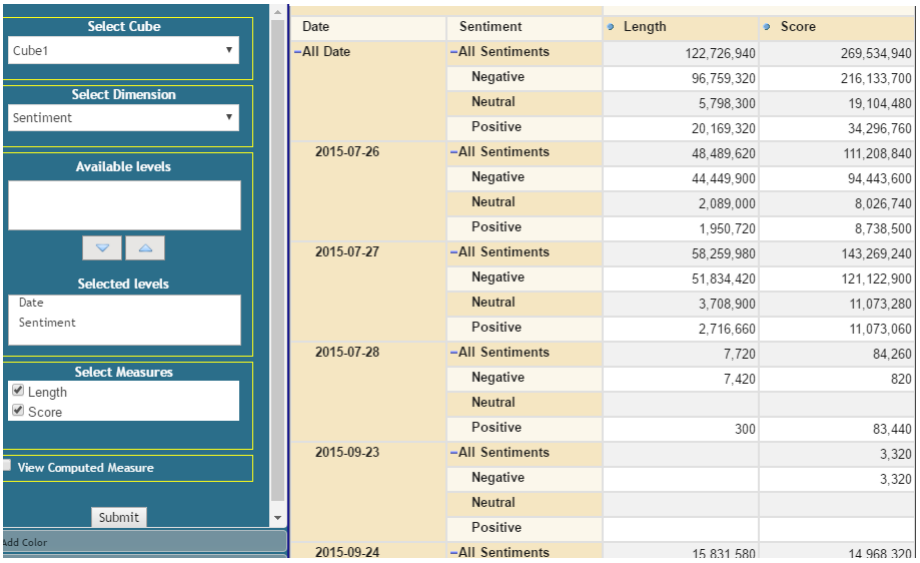
FIGURE 5.2: For all dates the measures like length and score is calculated

As I have selected my first cube where as a dimension i have select tweet creation date,that is created date and according to the following measures length of the tweet and score of each tweet is displayed. at the very first row, it calculates the total number of tweet length and the score.

Shows the length of the tweet and score with respect to the the sentiments and the created date of each tweet.

- Trend View:

The graphical bar represents the ups and down of trends depending on the measures, like for the first time first bar will come a as blue if there is any data. If there any at least 2 percent changes in compared to the previous month, it will come as green.If there is at least 2 percent decrement then the bar will come red. If the changes active using between plus minus 2 percent, then it will turn into blue. If the decrement comes, the trend will be red and down.Dimensional tables are used for the trend view where it the fact table has been changed. This helps to predict the future according to the historical data. Yellow bar shows the historical data, and related blue bar shows the future prediction value.



Date	Sentiment	Length	Score
All Date	All Sentiments	122,726,940	269,534,940
	Negative	96,759,320	216,133,700
	Neutral	5,798,300	19,104,480
	Positive	20,169,320	34,296,760
2015-07-26	All Sentiments	48,489,620	111,208,840
	Negative	44,449,900	94,443,600
	Neutral	2,089,000	8,026,740
	Positive	1,950,720	8,738,500
2015-07-27	All Sentiments	58,259,980	143,269,240
	Negative	51,834,420	121,122,900
	Neutral	3,708,900	11,073,280
	Positive	2,716,660	11,073,060
2015-07-28	All Sentiments	7,720	84,260
	Negative	7,420	820
	Neutral		
	Positive	300	83,440
2015-09-23	All Sentiments		3,320
	Negative		3,320
	Neutral		
	Positive		
2015-09-24	All Sentiments	15,831,580	14,968,320

FIGURE 5.3: Date and sentiments dimension with score measure

For a particular date the score of the positive ,negative and the neutral tweets can be accomplished.

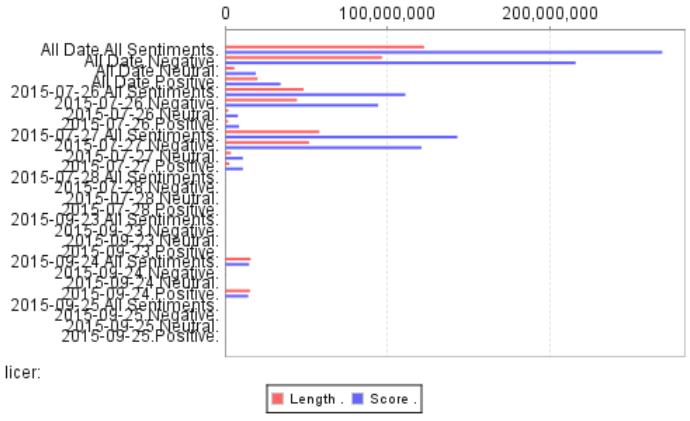


FIGURE 5.4: Graphical view of Fig 5.3

Shows the length of the tweet and score with respect to the the sentiments and the created date of each tweet.

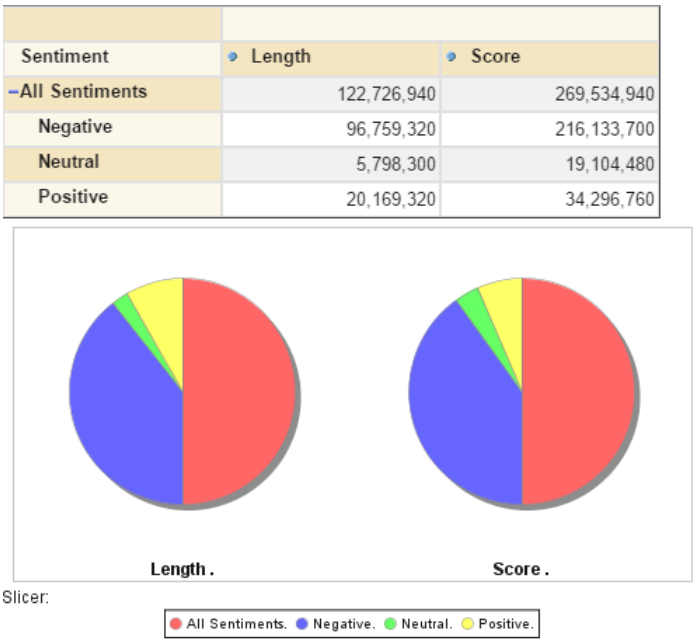


FIGURE 5.5: Sentiments with length and score measure

Depending on the sentiments tweets score are calculated.

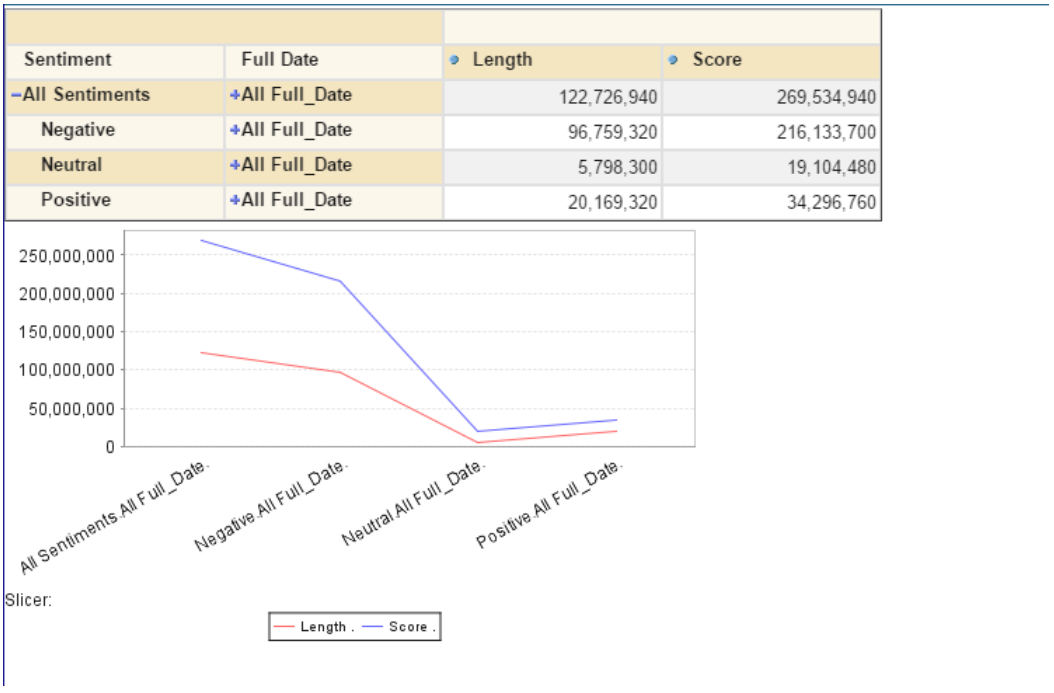


FIGURE 5.6: Sentiments with dates dimension and scores as a measure

Depending on the sentiments on a particular date tweets score are calculated.

Full Date	Length	Score
−All Full_Date	122,726,940	269,534,940
−2015-01-01		
−2015	122,726,940	269,534,940
−1		
−January		
−Q1		
Weekday	73,585,980	148,663,480
Weekend	49,140,960	120,871,460
+2		
+3		
+4		
+5		
+6		
+7	106,757,320	254,562,340
+8		
+9	15,969,620	14,972,600
+10		
+11		
+12		
+2015-01-02		

FIGURE 5.7: Dates will levels calculating score and length

Full date has labels where each date has year, then year has month number, month number has quadral and searches if it is weekday or weekend.

Sentiment	Time	Length	Score
−All Sentiments	+All Time	122,726,940	269,534,940
Negative	+All Time	96,759,320	216,133,700
Neutral	+All Time	5,798,300	19,104,480
Positive	+All Time	20,169,320	34,296,760

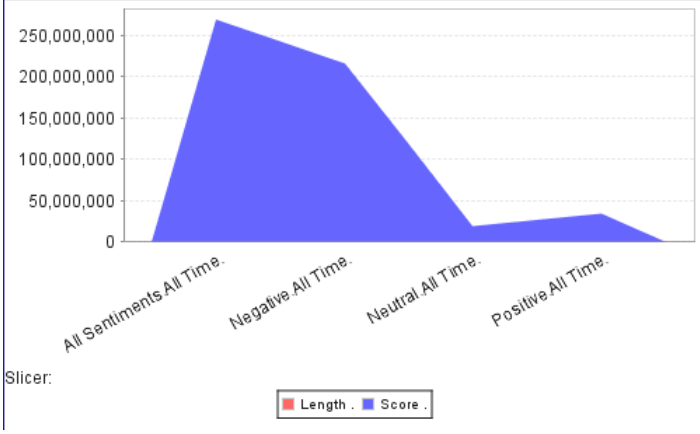


FIGURE 5.8: Sentiments with time dimensions calculating measures are length and score

Names	Network-Provider	Following Count	Followers Count
+All Names	−All Network_Provider	49,871,982	96,030,624
	DiGi	7,707	3,147
	Maxis	7,051,088	28,186,928
	MTN	42,653,896	67,756,003
	Orange	159,291	84,546

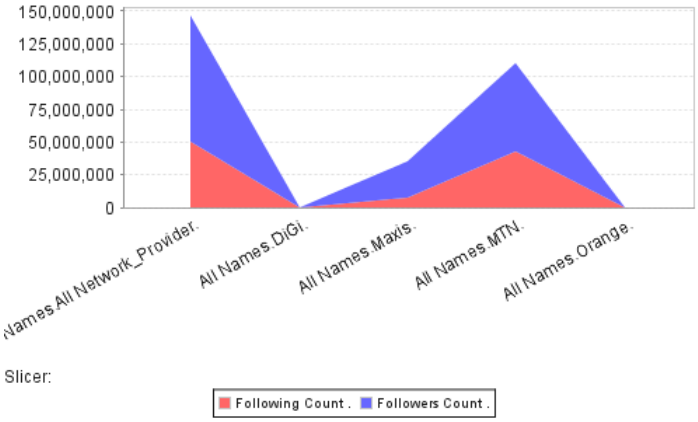


FIGURE 5.9: All users and network provider where measures ae following count, followers count

We have created another cube where for a particular network provider, how any people has followed measures is been calculated.

Network-Provider	Sentiment	Following Count	Followers Count
+All Network_Provider	-All Sentiments	49,871,982	96,030,624
	Negative	36,338,080	56,595,927
	Neutral	2,659,160	3,878,108
	Positive	10,874,742	35,556,589

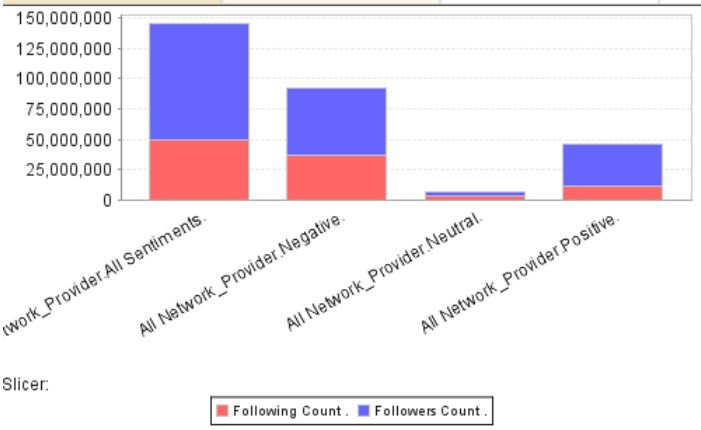


FIGURE 5.10: All the network provider with measures sentiments,following count, followers count

All the network provider with the sentiment dimensions are showing the aggregated count for the following count and the followers count. The graphical structure is also explained.

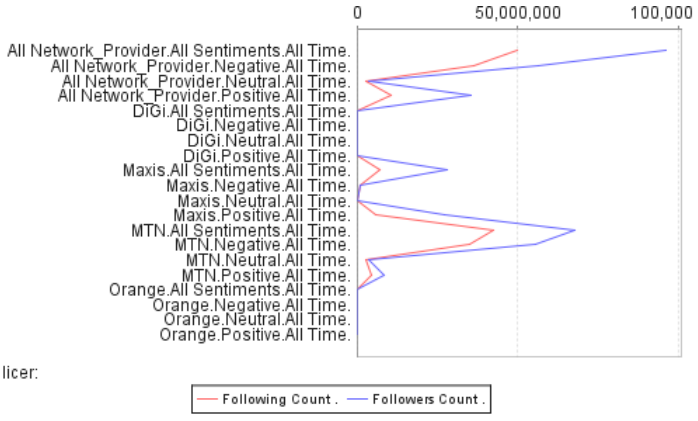


FIGURE 5.11: Graphical form of Fig:5.10

The graphical line chart is where all the sentiments and the network provider is plotted with time dimension and measures are following count and followers count.

Network-Provider	Sentiment	Time	Following Count	Followers Count
−All Network_Provider	−All Sentiments	+All Time	49,871,982	96,030,624
	Negative	+All Time	36,338,080	56,595,927
	Neutral	+All Time	2,659,160	3,878,108
	Positive	+All Time	10,874,742	35,556,589
DiGi	−All Sentiments	+All Time	7,707	3,147
	Negative	+All Time	1,437	1,506
	Neutral	+All Time		
	Positive	+All Time	6,270	1,641
Maxis	−All Sentiments	+All Time	7,051,088	28,186,928
	Negative	+All Time	987,888	994,664
	Neutral	+All Time	26,400	13,728
	Positive	+All Time	6,036,800	27,178,536
MTN	−All Sentiments	+All Time	42,653,896	67,756,003
	Negative	+All Time	35,275,528	55,533,867
	Neutral	+All Time	2,632,760	3,864,380
	Positive	+All Time	4,745,608	8,357,756
Orange	−All Sentiments	+All Time	159,291	84,546
	Negative	+All Time	73,227	65,890
	Neutral	+All Time		
	Positive	+All Time	86,064	18,656

FIGURE 5.12: network providers, sentiments and time dimension, following count and followers count measure

All the network provider with sentiments and time has plotted the aggregated value for measures following count and followers count.

Full Date	Network-Provider	Following Count	Followers Count
+All Full_Date	−All Network_Provider	49,871,982	96,030,624
	DiGi	7,707	3,147
	Maxis	7,051,088	28,186,928
	MTN	42,653,896	67,756,003
	Orange	159,291	84,546

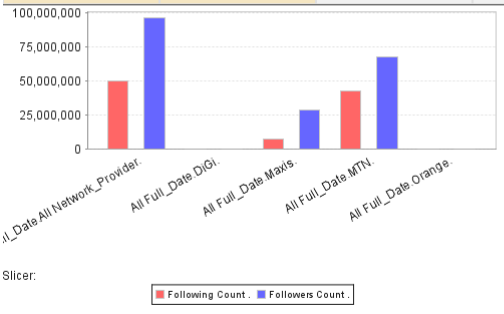


FIGURE 5.13: All dates and network provider, following count and followers count as the measures

For all specific date and for the particular network provider the following count and the followers count is aggregated.

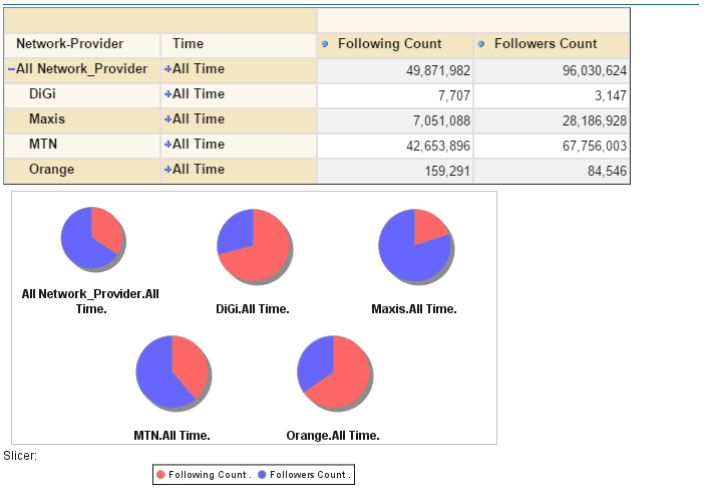


FIGURE 5.14: All network providers,all the time, with measure following count and followers count

Graphical representation of all the network provider with time and measures following count and followers count.

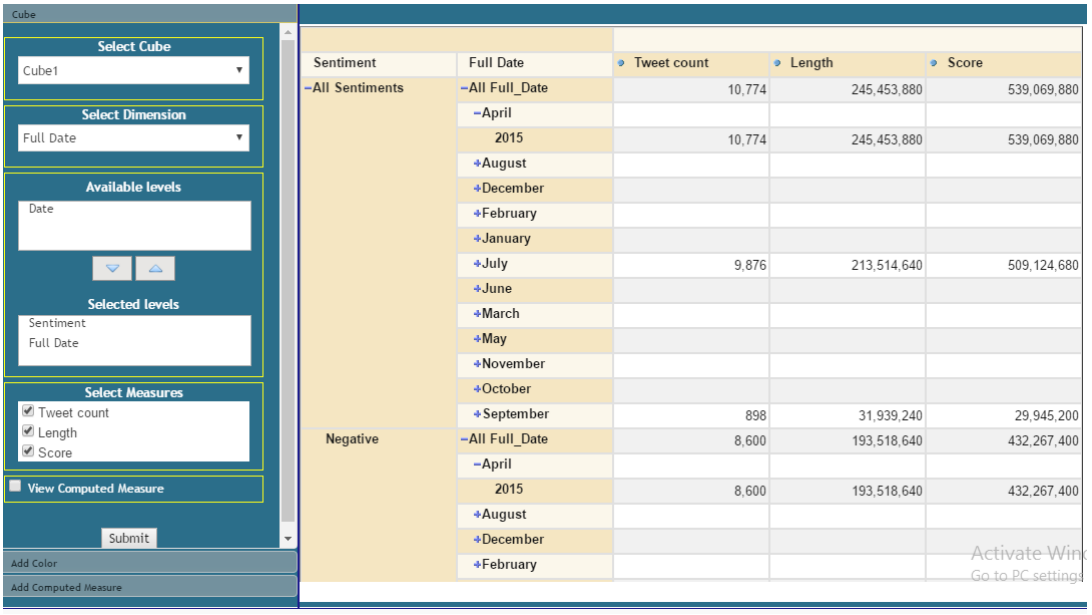


FIGURE 5.15: Sentiments and month name as in dimension and tweet count,following count and followers count in measures

Positive,negative and neutral sentiments with number of tweet count, length and score has been calculated.

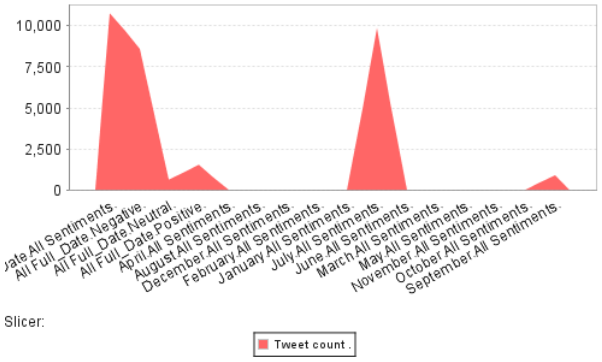


FIGURE 5.16: Graphical representation of FIG 15.5

With all sentiments and month the calculated tweet count is plotted.

Network-Provider	Sentiment	Tweet Count
−All Network_Provider	−All Sentiments	10,774
	Negative	8,600
	Neutral	636
	Positive	1,538
DiGi	−All Sentiments	6
	Negative	2
	Neutral	
	Positive	4
Maxis	−All Sentiments	884
	Negative	34
	Neutral	2
	Positive	848
MTN	−All Sentiments	9,876
	Negative	8,558
	Neutral	634
	Positive	684
Orange	−All Sentiments	8
	Negative	6
	Neutral	
	Positive	2

FIGURE 5.17: All network providers and sentiments with measure tweet count

Network providers and sentiments is the dimension which plots the tweet count as the measure.

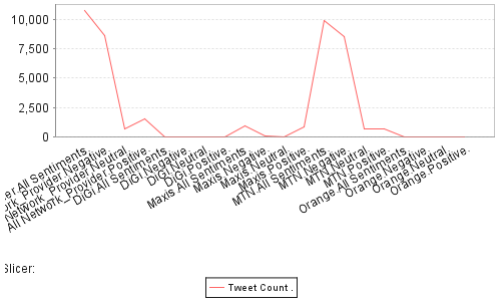


FIGURE 5.18: Graphical representation of FIG 15.7

Network provider and sentiments with the measure tweet count.

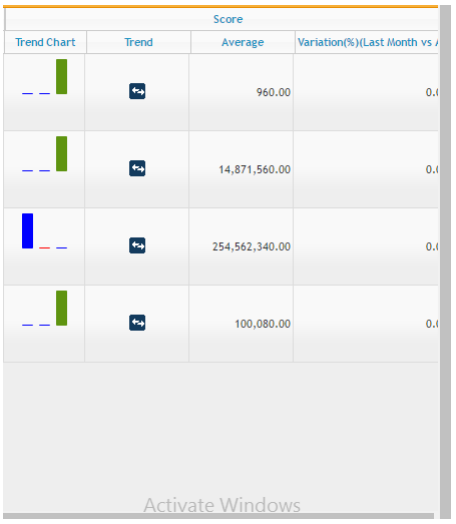


FIGURE 5.19: All the network providers trend value for 3 continuous month taking measures like score

One trend view application is shown where all the score is calculated for.

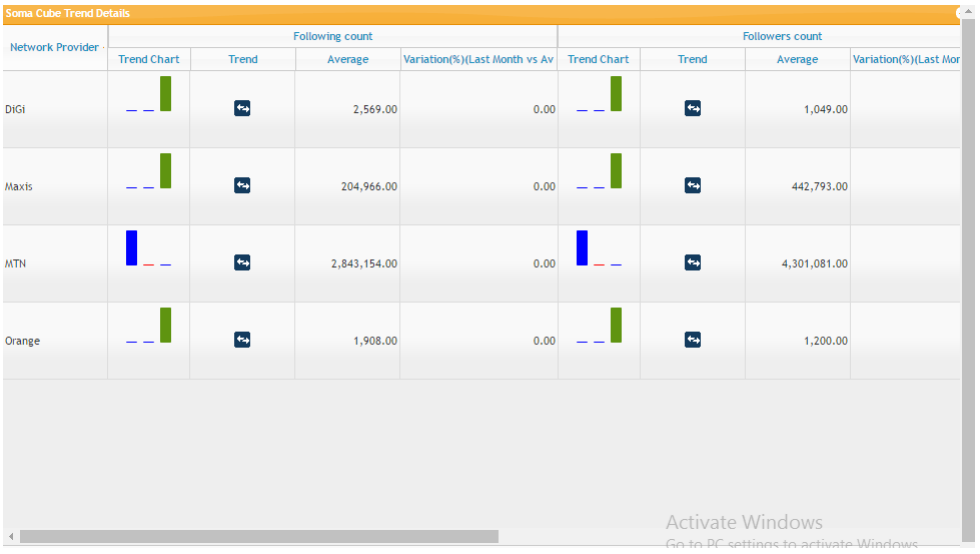


FIGURE 5.20: All the network providers trend value taking measures like following count and followers count

In the trend view application where all the network provider with following count and followers count has been calculated for 3 months time period.

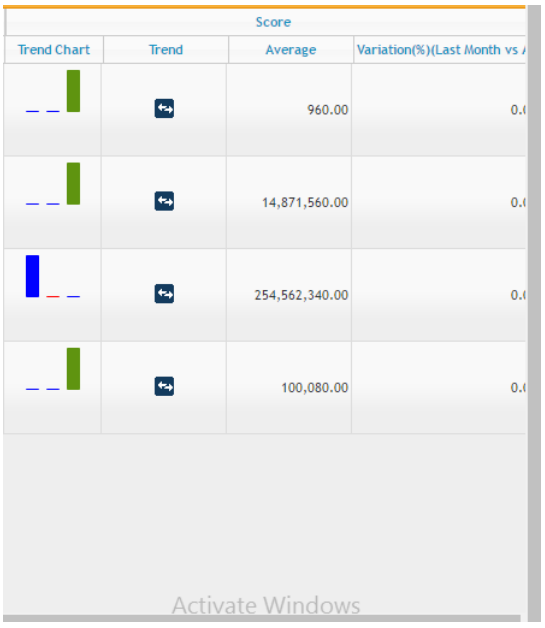


FIGURE 5.21: All the network providers trend value taking measures like tweet count

The trend view application for network providers with tweet count value for 3 months time.

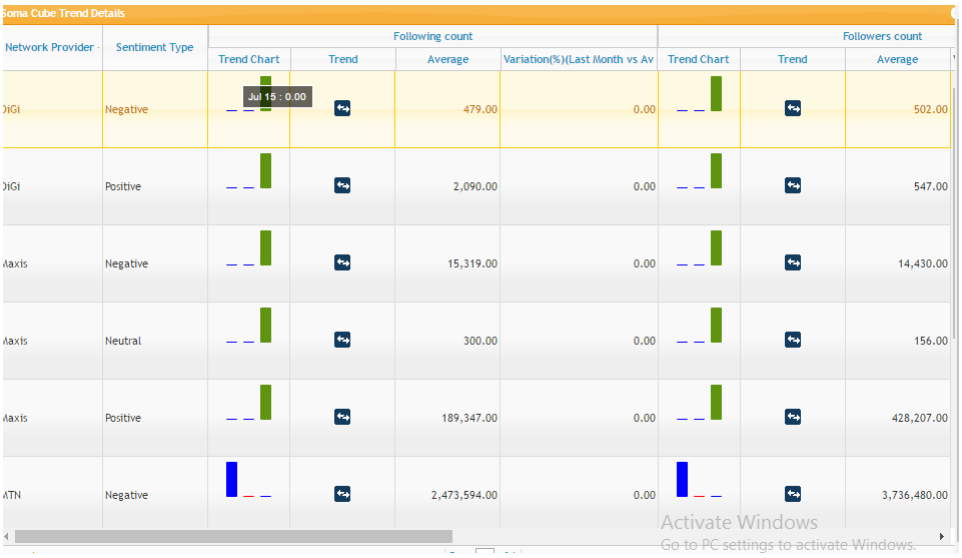


FIGURE 5.22: All the network providers and the sentiments, trend value for 3 continous month taking measures like following count and followers count

FIGURE 5.23: Network provider dimension with retweet count and score measures

All the network providers and the sentiments, trend value for 3 continuous month taking measures like retweet count and score

Historical data represents as orange and if there is sudden increase in prediction it will show green and if it is similar, it will come blue.

FIGURE 5.24: Predicted trend for all network provider with the measures

In trend view chart we have predicted the values of network provider with the measures following count and followers count after observing some months values.

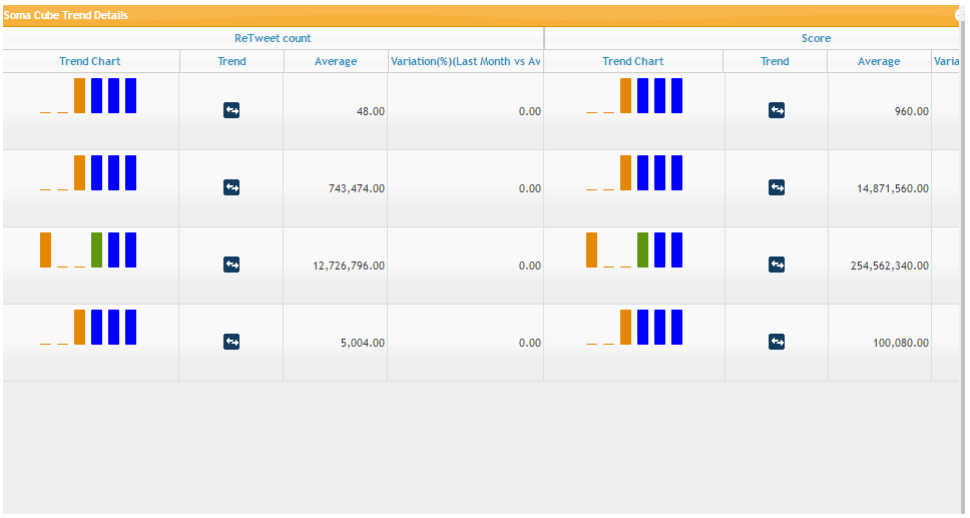


FIGURE 5.25: Predicted trend for all network Provider with the measures

In trend view chart we have predicted the values of network provider with the measures following count and retweet count and score after observing three months values.

FIGURE 5.26: Predicted trend for all network Provider and sentiments with the measures

In trend view chart we have predicted the values of network provider and sentiments with the measures following count after observing months.

FIGURE 5.27: Predicted trend for all network Provider with the measures

In trend view chart we have predicted the values of network provider and sentiments with the measures followers count and retweet count after observing 3 months values.

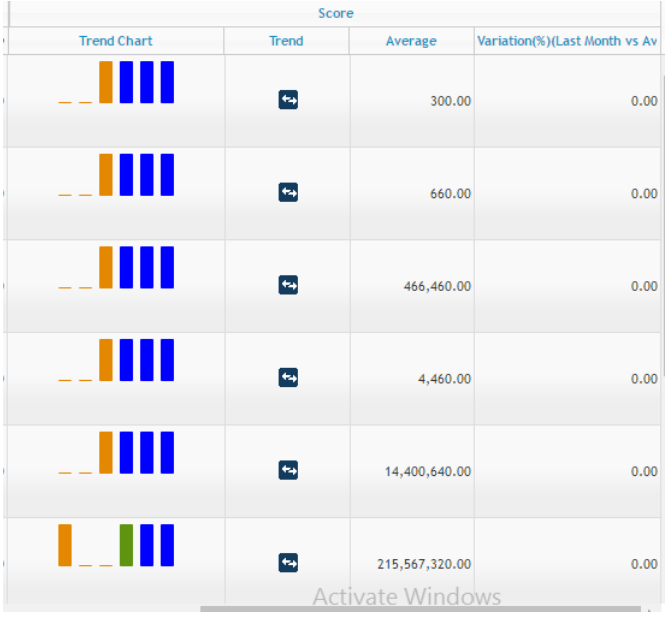


FIGURE 5.28: Predicted trend for all network Provider with the measures

In trend view chart we have predicted the values of network provider and sentiments with the measures score after observing continuous months values.

5.2 Conclusions

Data warehousing technology supports the study of huge data volumes caused by Twitter. We represent multi-layered system architecture which shows how to transform the data collected from twitter and the detailed explained process of each layer. OLAP cube is a method which represents the data in a multi-dimensional cube view. For many business area, while taking decisions this OLAP cube could be helpful. It can store a huge amount of data where at a time, a query can hit hundreds thousands of rows. Social media analytics gather the data from different sources and use them to fulfill organization's needs like understanding the market, customer needs etc. This analysis of data is done using sentiment analysis and represents the data in cube view where data is in many dimension with calculated measures and the trend view where it predicts the outcome based on the previous value over certain time period.

5.3 Scope for Future Work

For Naive Bayes classifier the training set I used is manually created. Twitter corpus can be created for using Naive Bayes classifiers which will give more accuracy of sentiments for each tweet. Another way visualization of data from cubes like dashboards can be implemented for business analysis.

Bibliography

[1] Narayanan, Vivek, Ishan Arora and Arjun Bhatia, "Fast and accurate sentiment classification using an enhanced Naive Bayes model," *Intelligent Data Engineering and Automated Learning–IDEAL.*, Springer Berlin Heidelberg, pp. 194-201, May 27, 2013.

[2] Holsapple, Clyde, Shih-Hui Hsiao and Ram Pakath, "Business Social Media Analytics: Definition, Benefits, and Challenges,",holsapple2014business, 2014.

[3] Agarwal and Apoorv "Sentiment analysis of twitter data," in *Proceedings of the workshop on languages in social media.*, Association for Computational Linguistics, pp. 30-38, Jun 23, 2011.

[4] Wiwatwattana and Nuwee, "X 3: A cube operator for xml olap," *Data Engineering, ICDE, IEEE 23rd International Conference on.*, IEEE, pp. 916-925, Apr 15, 2007.

[5] N. U. Rehman, S. Mansmann, A. Weiler and M. H. Scholl, "Building a Data Warehouse for Twitter Stream Exploration," *Advances in Social Networks Analysis and Mining (ASONAM).*, IEEE/ACM International Conference on, Istanbul, pp. 1341-1348, Aug 26, 2012.

[6] Liu and Xiong, "A text cube approach to human, social and cultural behavior in the twitter stream," *Social Computing, Behavioral-Cultural Modeling and Prediction.*, Springer Berlin Heidelberg, pp. 321-330, Apr 2, 2013.

[7] Mansmann and Svetlana et al. "Discovering OLAP dimensions in semi-structured data,"*Information Systems*, vol. 44, pp. 120-133, Aug 31, 2014.

[8] Rambocas, Meena and João Gama,*"Marketing research: The role of sentiment analysis"*.,Universidade do Porto, Faculdade de Economia do Porto, Apr, 2013.

[9] Liu and Bing,"Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies.*, vol. 5, no. 1, pp. 1-167, May 22, 2012.

[10] Prabowo, Rudy and Mike Thelwall, ”Sentiment analysis: A combined approach,” *Journal of Informetrics*., vol. 3, no. 2, pp. 143-157, Apr 30, 2009.

[11] Kiritchenko, Svetlana, Xiaodan Zhu, and Saif M. Mohammad, ”Sentiment analysis of short informal texts,”*Journal of Artificial Intelligence Research* , pp. 723-762, May, 2014.

[12] Godbole, Namrata, Manja Srinivasaiah, and Steven Skiena, ”Large-Scale Sentiment Analysis for News and Blogs,”*ICWSM*., vol. 7, no. 21, pp. 219-222, Mar 26, 2007.

[13] Thiel and Killian, ”Creating Usable Customer Intelligence from Social Media Data: Network Analytics meets Text Mining,” *KNIME. com Report*., Oct 4, 2012.

[14] Fan, Weiguo and Michael D. Gordon, ”The power of social media analytics,” *Communications of the ACM*., vol. 57, no. 6, pp. 74-81, Jun 1, 2014.

[15] Batrinca, Bogdan and Philip C. Treleaven, ”Social media analytics: a survey of techniques, tools and platforms,” *AI and SOCIETY*., vol. 30, no. 1, pp. 89-116,2015.

[16] Rehman and Nafees Ur, ”Building a data warehouse for twitter stream exploration,” *Proceedings of International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*., IEEE Computer Society, pp. 1341-1348, Aug 26, 2012.

[17] Kouloumpis, Efthymios, Theresa Wilson and Johanna D. Moore, ”Twitter sentiment analysis: The good the bad and the omg!,” *Icwsm*., vol. 11, pp. 538-541, Jul 17, 2011.

PUBLICATION DETAILS

Nilakshi Roy and Boppuru Rudra Prathap, “Data modeling and OLAP cube analysis on Twitter Stream: Survey Paper,” in *Proceedings of National Conference on Challenges and Opportunities in computer Engineering(NCCOCE’16),* organized by department of CSE, Christ University,Bengaluru and CSI, India., March 5 2016.

Appendix A

Appendix Source Code

Fetching raw tweets from twitter dynamically and storing in PostgreSQL

```
import twitter
import json
import psycopg2


api = twitter.Twitter(auth=twitter.OAuth('789506922-↩
    Q76Fuju6HXT5m4zkRd70TN4J4o9vTNnLUeXs1jDC', '↩
    UZAvAFrkD09QGJT9ZLrZ0B1kxi5bukHICKdpaxYDUzx5q', 'W94xcLqO4NhCrFKVSigwDk5yb', '↩
    U0LMMrj2eZOtxXr4zViYprO9nEuRJ8XEQt3iDD2vOqcKsIeZAW'))

conn = psycopg2.connect("dbname=testdb user=postgres password=postgres")
cur = conn.cursor()

cur.execute('SELECT text from word')
terms = cur.fetchall()

if terms:
    for t in terms:
        tweets = api.search.tweets(q="t", lang='en', count=100)

        for s in tweets['statuses']:
            cur.execute("SELECT * FROM tweet WHERE tid='%s'", (s['id'], ))

            # ignore the duplicate and RT-like tweets
            if cur.fetchone() is not None or s['text'].find('3g') >= 0:
                continue
            print s['id_str']
            print s['lang']
            print s['retweet_count']
            print s['favorite_count']
            print s['retweeted']
            print s['source']
            print s['user']['location']
```

```
            print s['user']['url']
            print s['features']['type']
            print s['user']['followers_count']
            print s['user']['listed_count']
            print s['user']['screen_name']
            print s['user']['description']
            print s['user']['statuses_count']
            print s['user']['following']
            print s['in_reply_to_user_id']
            print s['user']['friends_count']
            print s['user']['time_zone']
            print s['text']
            print s['user']['created_at']
            print s['user']['place_id']

        cur.execute("INSERT INTO tweet3 (id_str, lang, retweet_count, favorite_count,↩
     retweeted, source, location, followers_count, listed_count, screen_name, ↩
    description, statuses_count, following, in_reply_to_user_id, friends_count, ↩
    time_zone, text, created_at) VALUES (%s, %s, %s, %s, %s, %s, %s, %s, %s, %s,↩
     %s, %s, %s, %s, %s, %s, %s)", (s['id_str'],s['lang'],s['retweet_count'],s['↩
    favorite_count'],s['retweeted'],s['source'],s['user']['location'],s['user']['↩
    followers_count'],s['user']['listed_count'],s['user']['screen_name'],s['user']['↩
    description'],s['user']['statuses_count'],s['user']['following'],s['↩
    in_reply_to_user_id'],s['user']['friends_count'],s['user']['time_zone'],s['text↩
    '],s['created_at']))
        conn.commit()

cur.close()
conn.close()
```

Preprocessing of tweets for further analysis

```
import re
import sys
def processTweet(tweet):



    tweet = tweet.lower()

    tweet = re.sub('((www\.[^\s]+)|(https?://[^\s]+))','URL',tweet)

    tweet = re.sub('@[^\s]+','AT_USER',tweet)

    tweet = re.sub('[\s]+', ' ', tweet)

    tweet = re.sub(r'#([^\s]+)', r'\1', tweet)

    tweet = tweet.strip('\'"')
    return tweet
fp = open('twitterdata.csv', 'r')
line = fp.readline()
```

```
while line:
    processedTweet = processTweet(line)

    print processedTweet
    line = fp.readline()
    sys.stdout=open("preprocessed.txt","a")
fp.close()
```

Sentiment analysis on tweets

```
from __future__ import division
import urllib
import csv
from string import punctuation


files=['negative.txt','positive.txt','preprocess_orange.txt']


tweets = open("new5.txt").read()
tweets_list = tweets.split('\n')

pos_sent = open("positive.txt").read()
positive_words=pos_sent.split('\n')
positive_counts=[]

neg_sent = open('negative.txt').read()
negative_words=neg_sent.split('\n')
negative_counts=[]


for tweet in tweets_list:
    positive_counter=0
    negative_counter=0

    tweet_processed=tweet.lower()


    for p in list(punctuation):
        tweet_processed=tweet_processed.replace(p,'')

    words=tweet_processed.split(' ')
    word_count=len(words)
    for word in words:
        if word in positive_words:
            positive_counter=positive_counter+1
        elif word in negative_words:
            negative_counter=negative_counter+1

    positive_counts.append(positive_counter/word_count)
    negative_counts.append(negative_counter/word_count)

print len(positive_counts)
```

```
output=zip(tweets_list,positive_counts,negative_counts)

writer = csv.writer(open('new6.csv', 'wb'))
writer.writerows(output)
```

Appendix B

Appendix Paper Publication

Data modeling and OLAP cube analysis on Twitter Stream: Survey Paper

Nilakshi Roy
Christ University Faculty of Engineering
Bangalore, India
Email Id: roynilakshi91@gmail.com

Boppru Rudra Prathap
Christ University Faculty of Engineering
Bangalore, India
Email Id: boppuru.prathap@christuniversity.in

Abstract – Social media network are the platforms where user communicate, interact, share ideas, career interest, pictures, video etc. One of the famous micro blogging website is Twitter which has evolved a lot in the recent years. Twitter gives you the opportunity to analysis the human social behavior such as sentiments, influence, emotions etc. This paper covers the topic of the analysis of data ware house technology with the multidimensional data model, Extraction, transformation and loading (ETL) for stacking and renovating data from different sources and the design of OLAP (On-line Analytical Processing), a text cube approach to study the sentiments of the human based on the location, time and date. It also gives a benefit of visualizing the data in a simplest form of model that is Star schema. There data is structured in facts and dimensions. The main motive is to transform the tweet stream into structured data and represent them in multidimensional cube and schema model.. It also demonstrates them by executing some cube operations such as slicing, dicing, drill-up, drill-down and pivoting which can be used for solving a different analytical tasks. The benefits of the analysis platform is to visualize by creating a data warehouse model from the social network that is Twitter, dynamically elevating the defined dataset and enabling multidimensional cube analysis and doing cube operations.

Keywords– Data warehouse, sentiments, Star schema, Twitter, Microsoft SQL, OLAP, PostgreSQL.

I. INTRODUCTION

Outburst of social network activity in the latest years has led to generation of huge volume of user linked data which in turn gave the birth of the now a day's hot topic Social Media Analysis. Data warehousing is a technique for data warehouse which has become a good platform for most of the large companies. Companies, organizations and institutions all around the world have the opportunity to analyze public tweet stream to improve their marketing, customer services and also public relations from the learned knowledge. The increasing growth of web based social networks and the regular social activities results in huge social network data. Data can be of two type's one structured or unstructured like tweet, events, description, messages, brands, status, and comments. Data which has been stored in data warehouse as data cube must contain data of business process, customer reviews about some brand, distribution, sales, marketing etc. This data shows the customer pattern and trends, sentiments, business strategies and also many other characteristics. Data cube allows to aggregate numeric data and it is derived by measures and dimensions. So, the data which has been captured is valuable to the success of business and as well as to understand the human social behavior.

This work is to create a data warehouse for the public tweet stream for the purpose of its complete analysis. Twitter API allows the applications to get the dynamic data of tweet objects. Using API automatic live tweets for a particular topic such as 3g, 2g, price, rate-cutter can be fetched and then transform the data into structured one later load them into particular database to do some successive analysis. In this paper we will show the analysis of social media can be useful from this established and developed technology. We have introduced an OLAP cube approach for social media analysis, particularly sentiment analysis. The capabilities of OLAP has been extended by enrichment of the data set to find out new measures and dimensions for new data cubes and also supporting updated data as well as historical data. In data warehousing, data cube organize the data in multiple dimensions and several hierarchies for relevant data storing and visualizing from different perspectives. Data dimensional model, such as star schema, has been created using power query and power pivot in excel. A star schema model has some proper dimensions and a fact table. As facts contain the events and the dimensions holds the reference information about fact. This schema is used to support data warehouse and data marts OLAP cubes.

The paper is structured in such a way where it first describe brief about twitter, OLAP and star schema and the sentiment analysis of any text. Then we will be presenting Data warehouse architecture and describing each and every step in

detail. Later on we will be discussing about relational view and multidimensional view of tweet record and how many possible ways OLAP cubes can be made and the operations can be done on the cube. We will be concluding the paper by putting our own point of view about creating OLAP cubes and our future work.

A. *Twitter*

Twitter is the popular micro blogging network which supports real time information.

Originally it's been presented in 2006 as a platform for interacting by sending short messages through internet. In recent years it has gain the worldwide popularity and also got into broadcasting news such as an influential channel and also the means of exchanging real time information. Around 332 million active twitter-users, over 500 million of tweets generates regularly till May 2015. Twitter has the attention of political, commercial, research and other establishments by allowing tweet stream available to the public. Twitter provides a set of APIs for fetching the detail information about the users and their communications and also the twitter streaming API for data-intensive applications, the search for API for querying and clarifying the message content and the REST API for retrieving the principal primitives of twitter.

B. *Star Schema*

In data warehousing, star schema is the simplest form of any dimensional model, which has one or more fact tables which reference the dimension tables in order to structure the data. Fact table resides in the middle and it has surrounded by dimension tables. It gives a structure like a star. Each dimension is characterized in a single table and the primary key for each dimension is linked to a foreign key in the fact table. The dimensions in the star schema is related to all the measures of the fact table.

C. *Data Warehousing and OLAP*

Data warehousing and Online Analytical Processing useful in different components of systems which helps in decisions in support and business intelligence. It has been originated in the 90s where they have granted the permission for giving the access to the data to decision makers. Components that consist of for building these type of systems are databases and applications that offer the tools analysts which needs to support the decision making in the organizations. Data ware house is a specialization of database technology for integrating, accumulating, analyzing and also visualizing of data from different sources. It employed the multidimensional data model, which represents the data in a cube model which contains measures of interest. Data warehouse contains data that characterizes the business history of any company or organization. This ancient data is used for not only analysis; it also gives the provision of taking the business decisions at many levels something like strategic planning to routine evaluation of a discrete organizational unit.

OLAP tools give the benefits of creating online statistical reports by the means of query and the analysis of data warehouse information. Summaries are calculated using aggregate functions like AVG, SUM, MAX, MIN many more. There are such cube operations like slicing, dicing, drill-up, drill-down, roll-up, pivot etc. is used by user to explore multidimensional cubes.

Slicing is a procedure of selecting a subset of cube and taking a single value from the dimensions and also making new cubes from the rest of the dimensions. Dicing is an operation which creates a sub cube by permitting the analyst to pick some real values from the multiple dimensions. Drill-up and Drill-down where data can be concise and stretched from different ranges according to the user's choice. A roll-up operation which involves in briefing the data, along with the dimensions by using some formula. Lastly Pivot which allows the cube to rotate in space to see its different faces.

Applicability of data warehousing is constrained to some business scenarios. Data analysis has become crucial in variety of applications like non-business domains such as government, science, education, hospital, research, medicine etc.

D. *Linguistic analysis for sentiment*

Now a days the research are concerned on how sentiments are expressed in online reviews and news articles. This is a social and cultural way for human to express their emotions. Emotions reflect how an individual sees the world and also how he reacts to the world in a particular event. Year's back cognitive psychologists didn't take sentiments of human in considerations. Understanding individual or analyzing the behavior of a group of people helps to predict the future and also to get the decisions for any business process.

Emotion can be assumed not only non-verbally but also using text-based communication. In recent studies it has been found that human convey their emotions using text, blogs, emails, short messages and different textual conversation. If a person is already sad before any start of conversation, the other participant can understand his negative emotion in text based communication. Emotions can be classified as negative positive or neutral.

Suppose you have used some particular product and you wrote some reviews showing your emotions about that product brand in social websites. Company fetches all those reviews of the customers about its product in some particular area and takes proper business decisions.

II. DATA WAREHOUSE ARCHITECTURE

A data warehouse system captures and integrates the data from different sources for useful analysis and access. Data warehouse technology has become a leading solution for many broad data management. Figure 1 introduces the follow-on structure of the Twitter Data warehouse Implementation. The very first layer is data sources. Twitter has given the opportunity to collect the streamed tweet records using Twitter API. Twitter REST API provides programmatic access to the twitter and to read and write the tweet stream. It captures the author of the tweet, author details and also follower data and more. Twitter requires OAuth for all the API sessions that improve the platform between the users and the developers. OAuth secure and authenticate the platform and protect user privacy while improving tracking in the Twitter. The data which we have collected from twitter will be in json format. Twitter API keys are the unique keys which will be generated when you create an account in twitter. Go for the Twitter Application and fill a form and generate keys. Python has lots of libraries to support fetching data using API from Twitter, like Twython, Tweepy etc. Data sources also includes additional other sources, like event detection, graphical databases etc. which holds metadata of twitter.

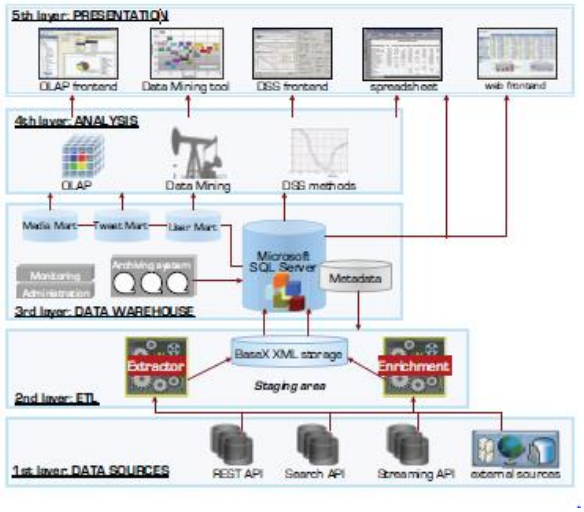


Fig.1. Twitter multi-layered architecture data warehouse system

The next layer is ETL (Extraction, Transformation and Loading) in Figure 3 which gathers the original data tweet stream and getting it prepared into the correct format with the final database and then feed forward the data as a dataset in the data warehouse. The data which we got from Twitter Streaming API has the JavaScript Object Notation format (json). When we fetch tweet stream more than 1000 tweets we get. Each tweet has 67 data fields in json format with high degree of the heterogeneity. Tweet record contains the tweet

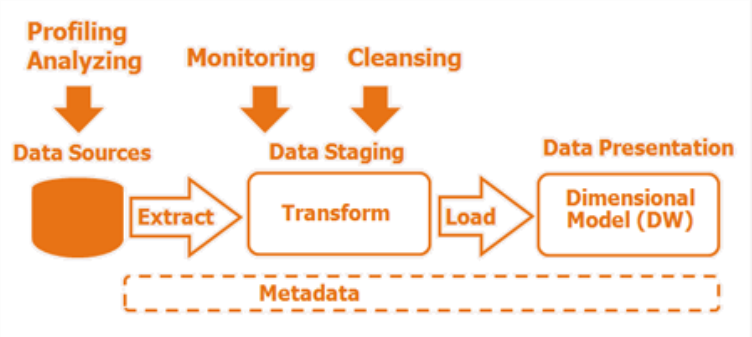


Fig.3. ETL stages

message plus all the comprehensive metadata about the users and its location. Almost 10% of public data by tweet streaming API covers 1million or more tweets/hour. This cause a weighty load of data in data warehouse.

The huge amount of data stream to convert into a single stream object into an XML structure and buffer in XML database Base X can be the solution.

Another solution is to fetch the tweet stream and while storing those, use filters which will only consider those particular Meta data as per as requirement and save it in database name Microsoft SQL or PostgreSQL. Create table in PostgreSQL or in SQL where the fetched data will be stored in an organized way.

Efficient BASEX storage can able to buffer high volume of streamed dataset which is impossible by buffering to the relational database. This loaded data is cleaned and the selective and needed dataset is stored in the Data warehouse so there will be much less load that the buffer. We use filters such as certain brand names, for some specific hours tweets etc.

We can also integrate data coming from different dataset. Suppose we have taken data from different APIs that combines into a single dataset. For this situation we use reverse geocoding to improve the graphic characteristic of every tweet for each city, country and continent. Then the ETL routine extracts this data to the star schema model of data warehouse. Microsoft SQL server maintenances powerful set of analysis services including OLAP and data mining.

The rest two upper layer of this architecture consists of the front end tools for analysis and visualization. First is the OLAP tools and next which is last is the end-user which is also the decision maker web based or in computer which generate from data's are reports, visual exploration or the executive dashboards etc. If we are using Microsoft SQL server then it would be easy to connect with the front-end tools of many open-source and commercial providers. If we use PostgreSQL it needs little effort to connect the database in Mondrian OLAP server to create OLAP cube using tomcat, Java and Jpivot.

The main challenging task is to implement a data warehouse for twitter where original data is stream delivered from twitter API transforms to structural organized multidimensional data set and later getting the multidimensional OLAP cubes.

III. MULTIDIMENSIONAL DATA MODEL OF TWITTER

It has become so important to visualize and analyze the following data model to get the knowledge that can be learnt from the data. This finds the relationships between the tweets of the user and the user itself. Users can be friends, followers who followed one user, or be referenced who had tagged one tweet or also can be the authors of the tweet or re-tweet other user messages. Another element is the timeline which describes the evolution or the arrangement of user and tweet objects. The underlying data model consists of the following objects:

- Status Object (tweets): It holds the text, the author and their metadata.
- User Object: It holds the user characteristics like screen name etc.
- Timelines: It provides an added view on the user's activity, such as the tweets which is authored by or mention a particular user who has tagged, status updates, followers, friendships relationships, re-tweets, friends count etc.

The above model is not necessary in OLAP cubes but can be taken as for multidimensional calculation. Each single data record in the stream holds Meta data, events along with the descriptions of user's profile. This dataset displays some characteristics of data warehousing such as

- Temporal by using time dimension
- Non-volatile that is no modifications in the existing data
- Measure-centric that is maintaining counters.

It's been observed that 67 fields of data in a single tweet stream which is a small number of attributes can describe some measures or for the analysis. This type of data structure can be taken from external sources or acquired by using various methods and functions from easy and simple computations to more complex techniques of knowledge discovery.

OLAP cube can be obtained from a twitter stream through a number of transformation steps.

A. Relational view of a record

The very first step is to get a structured model from the original tweet.

The main motive of this step is to identify the available entities, their value domains, constraints and the relationships between entities. UML notation is used for the structured elements. Figure2 shows the result of the relational mapping where a set of relations is linked by foreign key constraints.

The main class in this diagram is Tweet and User whereas rest of all the elements defines the relationship that relates either or both of them. User related information is the profile image, the location, the searches performed, notification received, users interaction. This interaction shows the statistics values as the counters on the followers and the following others, status updates, friends count. Tweet-related features comprise of location, the source of tweeting,

The other users mentioned, media embedded and also the count of re-tweeting and favoring the tweet. Relationship between the users and the tweet can be authoring or retweeting the message or being mentioned in the message.

Another table is sentiment. It has derived from the tweet reviews of the user. Suppose when we want to know the opinion of some specific brand of some specific area. Collecting all the reviews of the user, we can understand the sentiments whether it is positive, negative or neutral. Sentiment analysis is done using natural language tool kit (NLTK) or SVM (Support Vector Machine) etc. which all is predefined algorithms. NLTK has its own corpus name NLTK data which has predefined training set and also the corpus for movie reviews. While using NLTK we need to have training set and test set. Training set trains the test set data. Test records learn from the training set data and then predict. For sentiment analysis three training data set should be there one negative, one positive and one neutral which will classify the data in the test set after learning from the training set. Each review will be labeled as positive, negative and neutral.

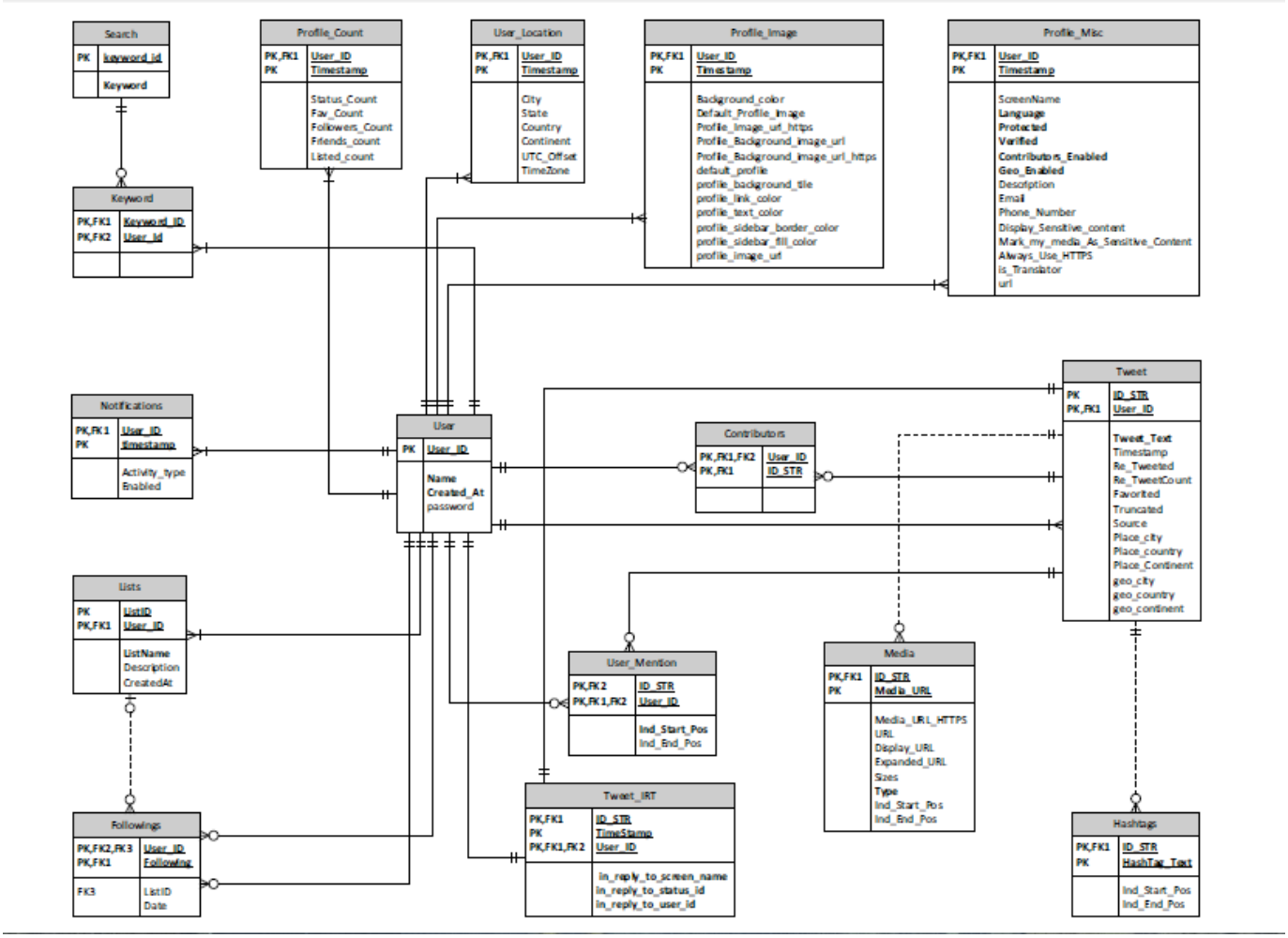


Fig.2. Conceptual model of the Twitter stream as a UML Diagram

B. Multidimensional view of a tweet record

Data in the data warehouse is well structured and organized. An entry in fact table consists of measurement, metrics of facts or events. Its location is the center of the star schema or the slow flake schema which is surrounded by the dimension tables. The fact consists of some measures like performance indicators with its whole description in the dimension table. Values in the dimensions are structured and roll-up and drill-down are the operations can be done. The representation of facts table and its associated dimension and the classification hierarchies is called *multidimensional data cube***.** In twitter star schema dimensions can be user details, date, location, product details, zone, and sentiments. Sentiments are the linguistic feature that can be added in the dimensions. Based on the Star Schema, data cube architecture can be designed to let the users to calculate the aggregated statistics of sentiment related measures along with the different dimensions like time and locations etc. Microsoft excel has given an opportunity to get the multidimensional data cube by adding some plugins.

Power Query is a data analysis feature available in Excel that allows you to define, combine and refine data. You only need to enable power query in excel to start. It helps in data discovery and a query tool that is good for shaping and mashing up data from different sources in a single workbook. It can import data, merge data, visualize power map and power view, share queries etc.

Power Pivot is an another features of Microsoft excel which allows to create tables and column properties, create relationships between them, add formulas with Data Analysis Expression(DAX), hierarchies, create key performance indicators(KPIs).

Import the data from the database to the excel sheet and from those data create table by the help of

power query. You can make table from selective columns. After making table you can visualize a star schema model using Power Pivot. The relation between the fact and the dimension tables are well cleared in Power Pivot.

VI. A TEXT CUBE APPROACH

OLAP cube is a structured form of the data which has been taken from twitter unstructured data. This type of representation helps in data analysis. Tweet sentiment adds an extra weightage while taking business decisions. In this data cube architecture, it allows the user to view the aggregate statistics of the measures of sentiments with the other dimensions like time, location, zone

Suppose, a data cube has created where product with day is in the horizontal dimensions and region is in the vertical dimension. Each of the measure of the tweet sentiment represents of the particular product on an exact day in a specific location. Measures can show the average of negative emotion for each day each for every region.

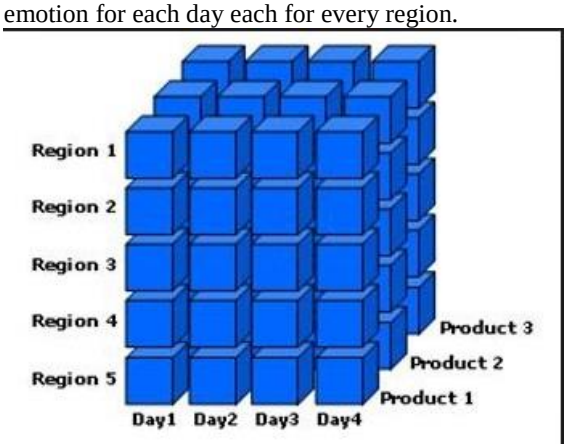


Fig.4. A data cube

Operations on cube such as slicing, dicing, drill-down, drill-up puts an additional advantage of viewing the cube and also helps to see the required values to make decisions. Slicing slices up the data cube like an example, if I want to measure the sentiment for a particular date and a particular location.

Dicing actually shows the limited portion of data like an example I want to see on last three days data along with the dimensions of region. Drill-up and drill-down is summarizing the data or detailing the data by navigating among the levels of data ranging. Likewise another operation is pivot. Pivoting rotate the cube as per as users comfort to see the data cube properly and to make business decisions.

OLAP cube can be created using many ways. One of the ways is Affective Process Cube. A cube has been represented with dimensions and measures. This cube schema model is done using XML. This schema represents different cubes with a logical model of various hierarchies, and members, this model can also map the logical model to the physical one. Schema helps to view the cube of tweets and perform the analysis on OLTP (Online Analytical processing Tool) where slicing and dicing and drilling operations are performed in cells.

Another way is SSAS (SQL Server Analysis Services) in Microsoft BI Tool, used for Online Analytical Processing. It uses Multi- dimensional expression (MDX). It is a query language that is used for querying a cube.

SSAS reads data from a multidimensional model, after that it shapes the schema in the BIDS (Business Intelligence Development Studio) and then generates dimensions, measures and cubes from that schema after fine alteration the cube. As per as the requirements the cube can be deployed. Statistics calculated measures and naming the sets can be done using MDX and user can later it browses the cube data in Excel as a client tool.

OLAP cube can be directly easily created from the database Microsoft SQL. For other databases like PostgreSQL that faces some complications while creating OLAP cubes. Java supports creating OLAP cubes using Jpivot and Tomcat application. Mondrian is an OLAP server which connects PostgreSQL.

VI. CONCLUSION AND FUTURE SCOPE

Data warehousing technology supports the study of huge data volumes caused by Twitter. Generally data warehouse stock traditional data where the source data needs to go through a number of transformations to some level of details. We represent multi-layered system architecture which shows how to transform the data collected from twitter using twitter API and the detailed explained process of each layer. OLAP cube is a method which stores data in used UML notation multidimensional form. It is used for data analysis and also for business intelligence. Here I have explained many ways of creating OLAP cube and also given a view on how OLAP operations can be helpful and used in business decisions making.

As a future enhancement, an OLAP cube can be implemented not only using twitter, it can be any social media data. Then analyzing the data and representing the data in multidimensional form which will help the decision makers to take appropriate decisions for company's betterment.

ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of the authors of the referred papers for their work and support of Christ University Faculty of Engineering for their constant support towards research.

REFERENCES.

[1] Nafees Ur Rehman, Andreas Weiler, and Marc H. Scholl "OLAPing Social Media: The case of Twitter", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), 2013.

[2] Jimmy Lin and Dmitriy Ryaboy," Scaling Big Data Mining Infrastructure:The Twitter Experience".

[3] Dragana Ćamilović, Dragana Bečejski-Vujaklija, and Nataša Gospić,"A Call Detail Records Data Mart: Data Modelling and OLAP Analysis".

[4] N. U. Rehman, S. Mansmann, A. Weiler, and M. H. Scholl, "Discovering dynamic classification hierarchies in olap dimensions," in *Proceedings of the 20th international conference on Foundations of Intelligent Systems* ser. ISMIS'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 425–434. [Online].

[5] E. Lo, B. Kao, W.-S. Ho, S. D. Lee, C. K. Chui, and D. W. Cheung, "Olap on sequence data," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008, pp. 649–660.

[6] C. X. Lin, B. Ding, J. Han, F. Zhu, and B. Zhao, "Text cube: Computing ir measures for multidimensional text database analysis," in *Data Mining 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 905–910.

[7] J. Han, Y. Chen, G. Dong, J. Pei, B. W. Wah, J. Wang, and Y. D. Cai, "Stream cube: An architecture for multi-dimensional analysis of data streams," *Distributed and Parallel Databases*, vol. 18, no. 2, pp. 173–197, 2005.

[8] M. Mathioudakis and N. Koudas, "Twittermonitor: trend detection over the twitter stream," in *Proceedings of the 2010 international conference on Management of data*. ACM, 2010, pp. 1155–1158.

[9] Lin, C., Ding, B., Han, J., Zhu, F., Zhao, B.: "Text Cube: Computing IR Measures for Multidimensional Text Database Analysis" In: Proc. 2008 Int. Conf. on Data Mining, Pisa, Italy,December 2008.

[10] Barbosa, L., and Feng, J. 2010. Robust sentiment detection on twitter from biased and noisy data. In *Proc. of Coling*. Bifet, A., and Frank, E. 2010. Sentiment knowledge discovery in twitter streaming data. In *Proc. of 13th International Conference on Discovery Science*.

[11] Esuli, A., and Sebastiani, F. "SentiWordNet: A publicly available lexical resource for opinion mining" In *Proceedings of LREC*., 2006.

[12] Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.:" Predicting elections with Twitter: What 140 characters reveal about political sentiment" In: International AAAI Conference on Weblogs and Social Media ,2010

[13] Dr. Sanjay Srivastava, Kaushal,Srivastava, and Akhil Sharma." Analysis of Telecommunication Database using Star Schema" International Journal of Latest Trends in Engineering and Technology (IJLTET)