



## Big Data Processing in Sentiment and Opinion Mining for Detecting Student Depression in E-Learning Using Rich Facebook Dataset Collection

<sup>1</sup>S. Priyanka\*, <sup>2</sup>M. Sivakumar

<sup>1</sup>Me Computer Science and Engineering, K.Ramakrishnan College of Technology,  
Under Anna University Chennai, Samayapuram, Trichy, India

<sup>2</sup>PHD research scholar, Department of Computer Science and Engineering, NIT, Trichy, India

**Abstract-** *The sentiment analysis holds the terms are sentiments (comments based on feeling) and opinion (comments based on thought or knowledge) to detect the depression only for student through the E-Learning system. Here introduce the facebook application interface as a “postbuk” for mining and authenticating the comments extraction from the facebook user and server. The mining datasets are pre-processed by NLP and classified by using support vector machine classification. The Cassandra a big data’s Hadoop framework is used for collects the large datasets from facebook to E-Learning system. Also used the Hadoop programming model, map-reduce for analysis the comments. The Social Data Analytics Tool (SODATO) that realizes the Extraction, Transformation and Loading (ETL) provisions social data analysis based on the context adaptive system. The sentiment analysis of text, apply this technique to the data analysis of big social data collected from Facebook page.*

**Keywords—** *PostBuk, Cassandra, Map Reduce, ETL, FBML*

### I. INTRODUCTION

The social communication is the great tool for exploring the thoughts and interests based on single topic. The sentiment analysis is combination of opinion, sentiment, emotions from the given texts. The depression detection for student, which could be used by online social network, is facebook. Here the facebook is one of the advantages for more formal and open mind conversation comments to analysis and beyond that the authentication satisfying is also most consideration. The datasets are considers as comments created from the user profiles, likes, share, mutual interest in communication based on education. Postbuk is a Face book application that provides the participating users with an interactive interface; to collect the information for sentiment analysis. This paper presents a new method for sentiment analysis in Face book that, starting from messages written by students, supports: (i) to extract information about the students education for sentiment polarity (positive, neutral or negative), as transmitted in the comments they write; and (ii) to model the users’ usual sentiment polarity and to detect significant emotional changes. The big data Hadoop framework is used for collecting the large amount of datasets in distributed environment. The use of computers in education has meant a great contribution for students. In order to provide, it is that Personalization necessary to store information about each student in what is called the student model. The specific information to be collected and stored depends on the goals of the adaptive learning system (e.g., preferences, learning styles, personality, emotional state, context, previous actions, and so on). Face book provides a distinctive advantage for this research: it is a network of friends. In Face book, the “wall” is the space where the users publish their own messages, contents and so on. Regarding text messages, there are several categories: status messages (each user writes them in his/her own wall), posts in others’ walls, and comments to either one’s or others’ publications. Therefore, we focused on social networks. There exist an increasing number of online social networks available through the Web. From these applications, “Face book” is the more popular around the world. The big data’s Hadoop Framework contains Hadoop Distributed File System (HDFS) and Hadoop Map Reduce. HDFS is a virtual file system that looks like any other file system, this file is split into many small files, and each of those files is replicated and stored on fault tolerance constraints. This paper is organized as follows. Section 2 presents the state of the art of the research areas related to our work. Section 3 de-scribes method for sentiment extraction. Section 4 presents the sentiments pre-processing. Section 5 presents sentiment analysis. Section 6 includes the sentiment classification. Section 7 presents system architecture in the context of E-learning and section 8 is opinion mining. Finally, the conclusions of the work done, along with some lines for future work, are presented in Section 9.

### II. RELATED WORK

When the datasets are large, some information fusion algorithms might not scale up well. For example, if an algorithm needs to load data into memory constantly, the program may run out of memory for large datasets. A simple and complete system for sentiment mining on large datasets using a Naïve Bayes Classifier with the Hadoop framework [8]. Facebook post identification (ID) is needed to allow the extraction of all the comments from the selected Facebook post (N. azmina m. zamani, siti z. z. abidin) [9]. To collect data, we created and registered a Facebook Connect

application, called *iFeel*. Allowing the app to require specific Facebook privileges allowed us to host the app from our Stanford.edu accounts, making it available to anyone, not just our friends, and allowing us to gather status updates quickly and efficiently [9]. The sentiment classification of user posts in Twitter during the Hurricane Sandy and visualize these sentiments on a geographical map centred around the hurricane. Then it show how users' sentiments change according not only to users' locations, but also based on the distance from the disaster [2]. Generally, IFrames offer a variety of manual configurations in regards with the FBML canvas pages in which most of the contents are automatically configured from Facebook [3]. Therefore, this raises the need of considering the different user characteristics in order to adapt the system according to the user needs and other relevant aspects. [7] Finally, in recent years, due to the increasing amount of information delivered through social networks, many researches are focusing on applying sentiment analysis to these data [1]. In this direction, we take the simplified definition of sentiment for analysis the retrieved datasets. [1]

### III. SENTIMENT EXTRACTION

To identify the sentiments, we have to extract the datasets from social media that is Facebook. Here we follow some techniques and methods to extract the comments in sentiments. So here describes the following methods are REST protocol, PostBuk API, Datasets collections, FBML, ETL tool-PENTAHO DATA INTEGRATION.

#### 3.1 REST protocol

REST stands for **R**epresentational **S**tate **T**ransfer. (It is sometimes spelled "Rest"). The web server is send the request to FB server to retrieve the dataset based on sentiments in education. The request and response From FB sever is based on the fully http protocol. This support the stateless, the request and response is independent in the whole system. Here the web server is act as REST CLIENT and FB sever is act as REST SERVER. Also it's a client server and cache communication protocol is more advantage for fetching and analysis. The rest is an architecture style for designing networked application. Also light weighted alternative to mechanism like remote procedure call and web services like WSDL. The Facebook API is a platform for building applications that are available to the members of the social network of Facebook. Fig1 describes the API allows applications to use the social connections and profile information to make applications more involving, and to publish activities to the news feed and profile pages of Facebook, subject to individual user's privacy settings. With the API, users can add social context to their applications by utilizing profile, friend, Page, group, photo, and event data. The API uses Restful protocol and responses are in JSON format. [7] We'll get the UID associated with the current Facebook API session key by calling the API method `users.getLoggedInUser`. (// Last parameter specifies that this API call's result.// should be returned to us as a Long value.)

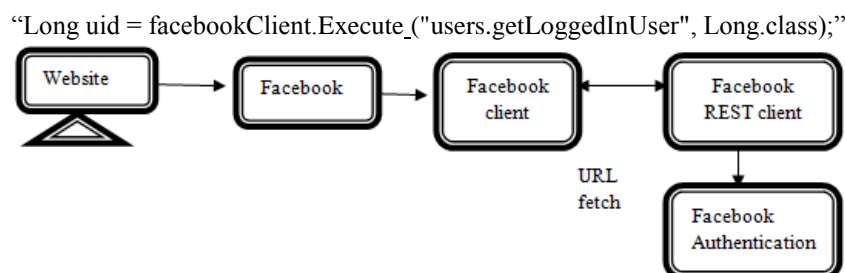


Figure1: REST protocol

In addition to LegacyFacebookClient, RestFB provides default implementations for Web Requestor and JsonMapper, two components that DefaultFacebookClient depends on to do its work. These dependencies are designed to allow for straightforward sub classing (if you only want to replace a little bit of functionality) and simple custom implementations (if you require full control). This comes in handy when unit testing - for example, you can write your own Web Requestor implementation that simulates a Facebook API endpoint response. You can drop in custom data designed to exercise your application's Facebook integration or simulate error conditions to make sure you're handling them properly. [7]

```

“LegacyFacebookClient facebookClient = new DefaultLegacyFacebookClient(MY_ACCESS_TOKEN,”
// A one-off DefaultWebRequestor for testing that returns a hardcoded JSON
// list of numbers instead of hitting the Facebook API endpoint URL
“ new DefaultWebRequestor() {
@Override
public Response executePost(String url, String parameters) throws IOException {
return new Response(URLConnection.HTTP_OK, "[123,456,789]”);
}
}, new DefaultJsonMapper());”
// Make an API request using the mocked Web Requestor
“List<Integer> numbers = facebookClient.executeForList("ignored", Integer.class); “
// Make sure we got what we were expecting
“assert numbers.size () == 3;
assert numbers.get (1) == 456; “
    
```

### 3.2 Postbuk API

According to the Facebook platform every web Facebook-connected application has to be authenticated and authorized from the Facebook user. Therefore, Facebook, in order to protect the privacy of the users who have not explicitly authorized a certain application, they are obliged to give access or leave the application. By default, the acceptance of any other application only permits the access of basic profile information of Users, such as their names, profile picture, the list of their Facebook friends and any other information they have shared with. As in this project we aimed at investigating the Facebook activity between the participants and their friends, it was necessary to have more detailed information about their behaviour in education into the Facebook world.[3] Fig2 describes, for this purpose, we needed to ask for additional, special extended permissions. For instance, permissions to access all the profile information of the application users and their Facebook friends, such as their email, friend list, comments and likes about their uploaded videos, photos and albums that had been posted from any other Facebook user. In order to collect the user data, any Facebook application has to ask for the user's permission, following a strict policy oriented to-wards protecting the user's privacy. The user has to give different permissions to access to each specific piece of information or to allow the application to use particular methods to access to this information. Therefore, the data that can be accessed by a given application depends on the permissions that the user granted to that application. For this reason, the first step once any user accesses PostBuk for the first time is to ask him/her for the following permissions:

Offline\_access: Permission to access the user data even if he/she has no opened session in Facebook at that time.

Read\_stream: Permission to read the user wall.

User\_about\_me: Permission to read basic user data (gender, birthday, languages, etc.).



Fig2 POSTBUK Page

### 3.3 Datasets Acquisition

The datasets in existing we are used here KONECT (Koblenz Network Collection). Datasets in KONECT represent networks, i.e., a set of nodes connected by links. Networks can be classified as (directed/undirected/bipartite), by their edge weight types and multiplicities, by the presence of metadata such as timestamps and node labels, and by the types of objects represented by nodes and links. Fig 3 denote in order to provide a unified view on such network datasets, and to allow the application of network analysis methods across disciplines, the KONECT project defines comprehensive network taxonomy and provides a consistent access to be Facebook network datasets. To validate this approach on real-world data from the Web, KONECT also provides a large number (210+) of network datasets of different types and different application areas. KONECT, the Koblenz Network Collection, contains 214 network datasets as of October 2014. The datasets we are extracted from the datasets collections. For this purpose, it was necessary to store their e-mail addresses.

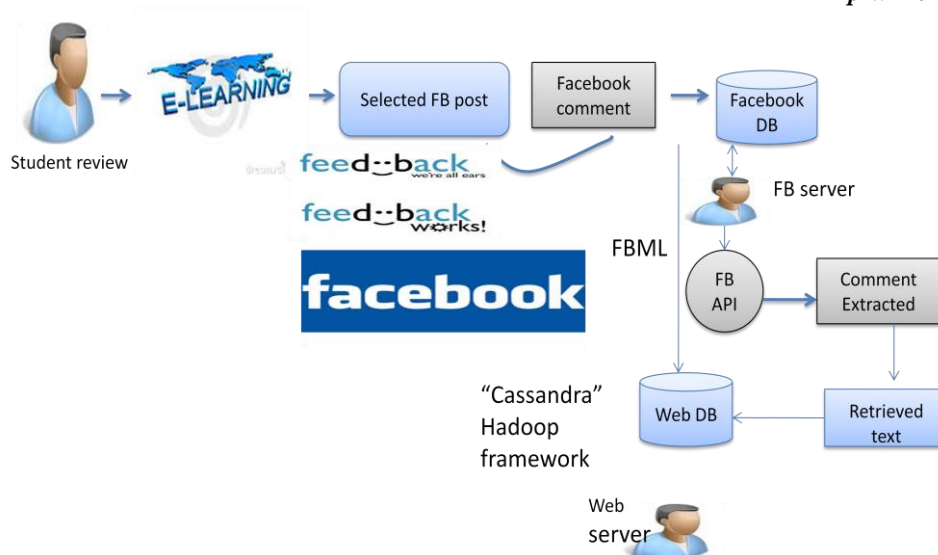


Fig 3 Module diagram for datasets acquisition

### 3.4 FBML

FBML (Facebook markup language) can be an ideal tool for organizing your Facebook applications. One FBML tag on a page takes no HTTP requests off your servers in order to render, whereas a simple API call for the same information could take one HTTP request to authenticate with Facebook and another to retrieve the information desired. FBML is a tool, a resource, for you as a developer to make your work in the Facebook API more efficient and reduced the API call. [3]. When the user requests (1) a FBML page Facebook does not send back a response immediately; instead, it sends (2) an HTTP POST to a call-back URL in the application server. Afterwards, Facebook expects from the application server to return (5) FBML, in order to convert it into HTML and finally, send (6) it back to the user's browser. Sometimes, the FBML pages need to make intermediate API calls (3), (4) between Facebook and the application server, adding some extra delay until the final response [3]. (Fig 4) However, when the application needs to show Facebook data such as names in FBML, it can avoid making calls to the Facebook API, using tags to reference the data directly moreover, there are Facebook servers directly peered with some of the largest hosting companies that serve application pages. Thus, the best choice was to enable only the FBML tags integrating the functions of the Facebook API. [3]. It describes the whole process with http protocol process.

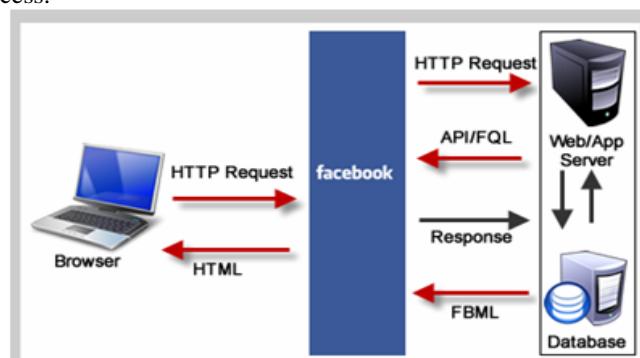


Fig 4 FBML process

### 3.5 ETL Tool -Pentaho data integration

Pentaho Data Integration (PDI) is a powerful extract, transform, and load (ETL) solution that uses an Innovative metadata-driven approach. It includes an easy to use, graphical design environment for building ETL jobs and transformations, resulting in faster development, lower maintenance costs, interactive debugging, and simplified deployment. **Spoon** is the design interface for building ETL jobs and transformations. Spoon provides a drag and drop interface allowing you to graphically describe what you want to take place in your transformations which can then be executed locally within Spoon, on a dedicated Data Integration Server, or a cluster of servers.

PDI is unique in that it can execute both outside of a Hadoop cluster and within the nodes of a Hadoop cluster. From outside a Hadoop cluster, PDI can extract data from or load data into Hadoop HDFS, Hive and HBase. When executed within the Hadoop cluster, PDI transformations can be used as Mapper and/or Reducer tasks, allowing PDI with Pentaho Map Reduce to be used as visual programming tool for Map Reduce (Fig5). By using this extract the comments from facebook directly.

```

Hadoop fs -mkdir /user/pdi/weblogs
Hadoop fs -mkdir /user/pdi/weblogs/parse
Hadoop fs -put weblogs_parse.txt /user/pdi/weblogs/parse/part-00000
    
```

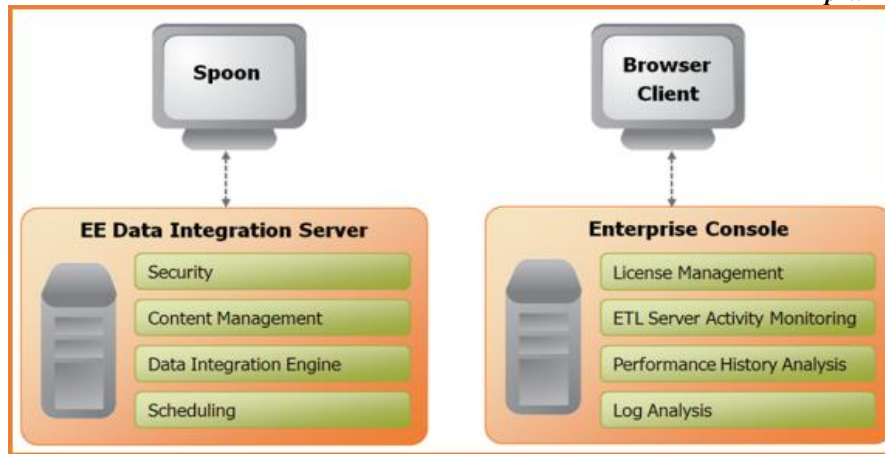


Fig 5 System Architecture for Pentaho

#### IV. SENTIMENT PRE-PROCESSING

The pre-processing is mainly done for eliminating the noise from extracted datasets. Before we had done the analysis, the NLP pre-processing is used. Then interquartile range pre-processing is used for the datasets pre-processing. It determines the whole outlier and extreme value range accurate which present in the datasets extracted. Fig 6 describes, the following techniques are stop words, stemming, part of speech (POS), and remove the repeated letter, interjection detection, tokenization, spell check also done. This process makes the datasets more accurate range of the analysis. [1]

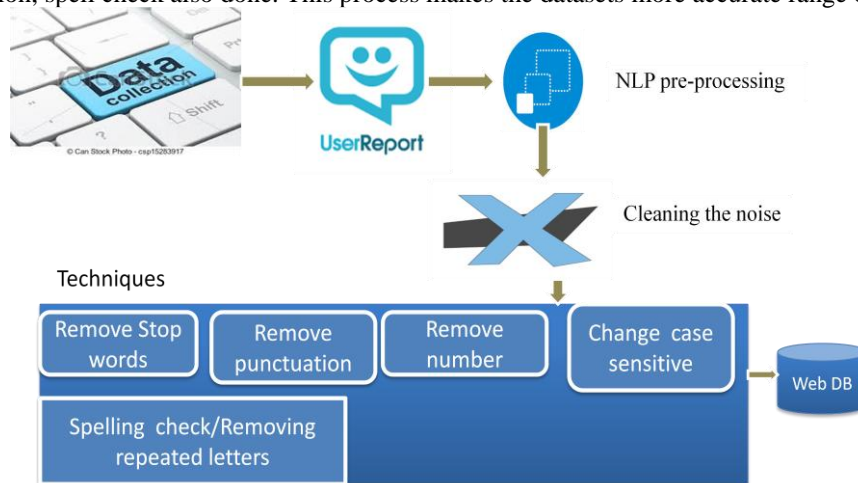


Fig 6 Module diagram for dataset Pre-processing

#### V. SENTIMENT ANALYSIS

The pre-processed datasets are collected and there is a possible of many replications of the comments given by the users. So here used the datasets analysis process with map-reduce and big database Cassandra.

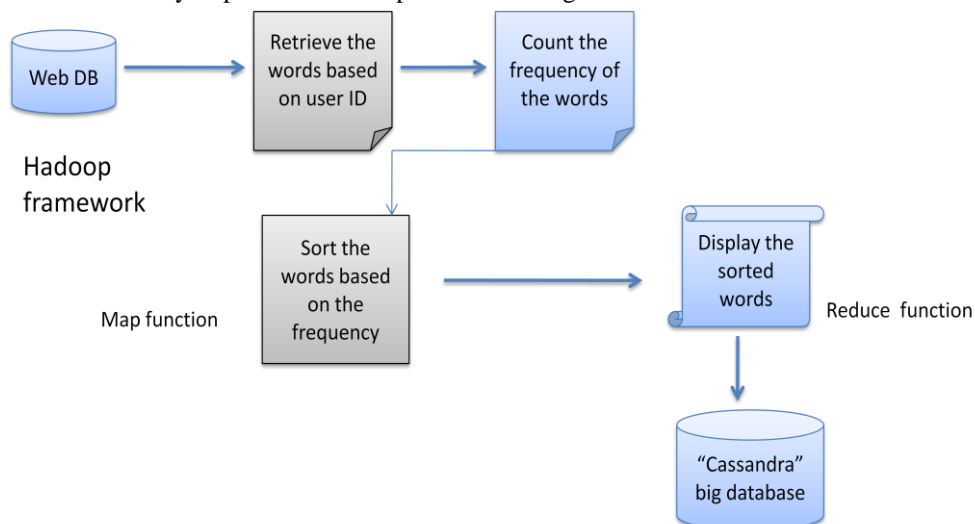


Fig 7 Module diagram for dataset Analysis



## 5.1 Map-Reduce

Map Reduce is the software programming framework in the Hadoop stack that simplifies processing of big data sets. A Map Reduce job consists of at least a map function and a reduce function, called Mapper and reducer respectively. The Mapper takes as input a pair of key/value and produces a set of key/value pairs. Fig 7, 8 represents all key/value pairs are sorted by their keys and sent to different reducers according to the key. Each reducer receives a key and a set of values that has the same key. This makes Map Reduce an excellent tool for computations that need Sorting or counting. The map and reduce functions are left to the user to implement their desired functionalities to process each key/value pair. Hadoop Map Reduce (Hadoop Map/Reduce) is a software framework for distributed processing of large data sets on compute clusters of commodity hardware. According to The Apache Software Foundation, the primary objective of Map/Reduce is to split the input data set into independent chunks that are processed in a completely parallel manner. The Hadoop Map Reduce framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically, both the input and the output of the job are stored in a file system. [11]



Fig 8 Map-Reduce

## 5.2 Cassandra

Infact, its open-source nature has given birth to a huge Cassandra community where like-minded people share their views, queries, suggestions related to Big Data. Cassandra can be integrated with other Apache open-source projects like Hadoop (with the help of Map Reduce). Cassandra follows a peer-to-peer architecture, instead of master-slave architecture. Hence, there is no single point of failure in Cassandra. Moreover, any number of servers/nodes can be added to any Cassandra cluster in any of the data centers. As all the machines are at equal level, any server can entertain request from any client. Undoubtedly, with its robust architecture and exceptional characteristics, Cassandra has raised the bar far above than other databases. Interestingly, any number of nodes can be added or deleted in Cassandra cluster without much disturbance (Table I describes). In a Cassandra cluster, each row is replicated based on the row key. You can set the number of replicas you want to create. Just like scaling, data replication can also happen across multiple data centers. This further leads to high level back-up and recovery competencies in Cassandra. Thus, Cassandra is used by those organizations that deal with huge amount of data every day and at the same time cannot afford to lose such data. Cassandra has a very high-level data model – this is column-oriented. It means, Cassandra stores columns based on the column names, leading to very quick slicing. Unlike traditional databases, where column names only consist of metadata, in Cassandra column names can also consist of the actual data. Thus, Cassandra rows can consist of masses of columns, in contrast to a relational database that consists of a few numbers of columns. Cassandra is endowed with a rich data model. Eventual consistency makes sure that the client is approved as soon as the cluster accepts the write. Whereas, Strong consistency means that any update is broadcasted to all machines or all the nodes where the particular data is situated. You also have the freedom to blend both eventual and strong consistency. For instance, you can go for eventual consistency in case of remote data centers where latency is quite high and go for Strong consistency for local data centers where latency is low. Cassandra there is no need to show all the columns needed by your application at the surface as each row is not expected to have the same set of columns. It is because of the above reasons; Cassandra is in great demand among several companies, where MySQL is getting replaced by NoSQL databases. A database that was initially created to solve the inbox search issues at Facebook has come a long way to solve Big Data problems. Let's look at some of the companies like Facebook, eBay, used Cassandra.

Table I Database Used In Social Media

Candidate	Usage	Tools
Facebook	Email search system containing 60TB+ and over 100m mailboxes	Cassandra
Google	Used for generating and modifying most of their data	BigTable [proprietary] (Equivalent to open source Hadoop Hbase)
LinkedIn	Handle hundreds of millions of reads and writes per day from over 400ms to under 10ms	Voltemort

## VI. SENTIMENT CLASSIFICATION

SVM is a machine learning classifier widely used for text categorization. The review text to be classified is converted into word vectors. SVM constructs a hyper plane using these vectors which separates data instances of one class from another. SVM finds this hyper-plane using training instances also called support vectors. In the binary categorization of text, the hyper-plane which classifies document  $d_j$  as  $c_j \in \{1, -1\}$  can be represented by weight vector of  $\vec{w}$  [11]. (Fig 9)

$$\vec{w} := \sum \alpha_j \vec{d}_j, \alpha_j \geq 0;$$

Where  $\alpha_j$  is a multiplier and for  $\vec{d}_j$  that  $\alpha_j$  are greater than zero are support vectors [11]. Test instance is classified by determining which side of  $\vec{w}$ 's hyper-plane they fall on. [1] Support Vector Machine (SVM) is a machine learning tool that is based on the idea of large margin data classification. The tool has strong Classification algorithms based on it give good generalization performance. Standard Implementations are providing good classification accuracy. They typically need large number of support vectors. Hence the training as well as the classification times is high. This algorithm selects new support vectors from a random sample based on generalization ability.

Experimental results done on real-world large datasets show that these methods help to reduce the storage cost, produce comparable classification accuracy with existing works and result in reduction of support vectors thereby reducing the inference time. Binary pattern recognition involves constructing a decision rule to classify examples into one of two classes based on a training set of examples whose classification is known a priori. Support Vector Machines (SVMs) construct a decision surface in the feature space that bisects the two categories and maximizes the margin of separation between two classes of points. This decision surface can then be used as a basis for classifying points of unknown class. [1]

Algorithm for SVM trained module:

Candidate SV= {closest pair from opposite classes}

While there are violating points do

Find a violator

Candidate SV = candidate SV  $\cup$  Violator

If any  $\alpha p < 0$  due to addition of c to S then

Candidate SV = candidate SV  $\setminus$  p

Repeat till all such points are pruned

End if

End while

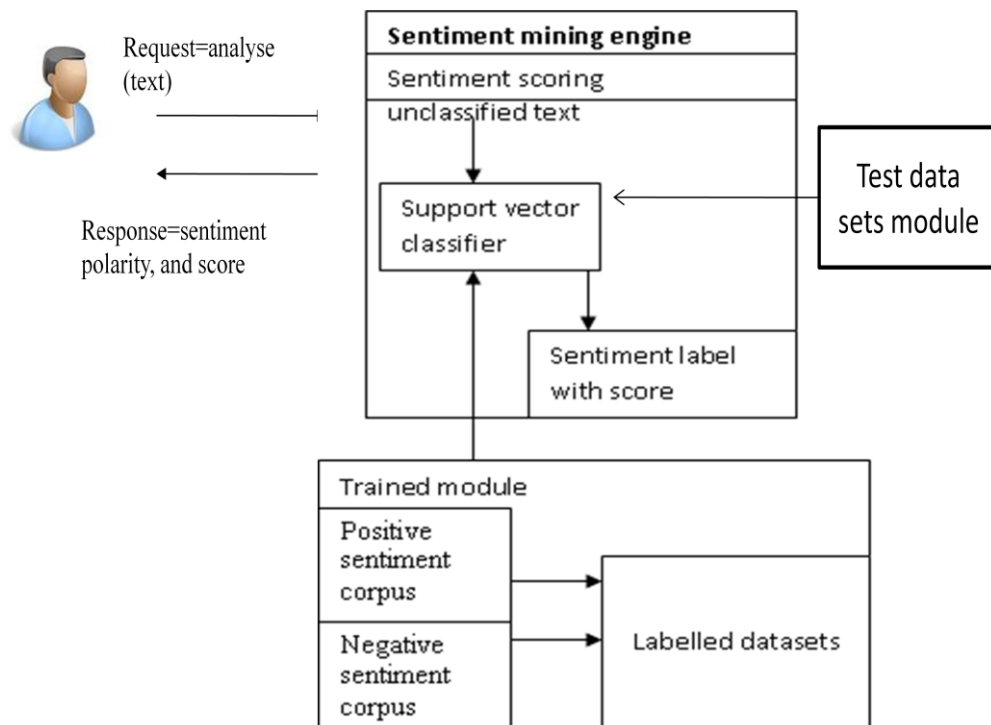


Fig 9Module diagram for datasets classification

## VII. SYSTEM ARCHITECTURE

In the system design phase, the server, program and database are considered. Firstly, the web server is needed for data pre-processing in the comments extraction. Then, related words are collected and stored into a database handled is big data's Hadoop and map-reduce for sentiment analysis. Hence, the design involves several components as shown in Fig10. [2]

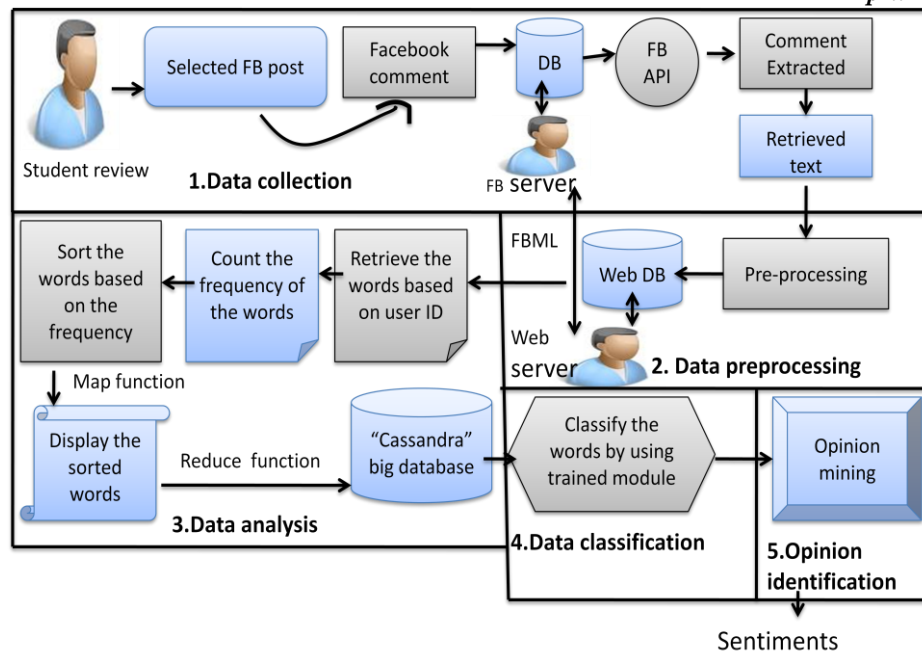


Fig10 System Design

## VIII. OPINION MINING

Opinion Mining also called sentiment analysis is a Process of finding user's opinion towards a topic. Opinion mining concludes whether user's view is positive, negative, or neutral about product, topic, event etc. Opinion mining involves analyzing user's opinion, attitude, and emotion towards particular topic. This consists of first categories text into subjective and objective information, and then finding polarity in subjective text. Opinion Mining can be performed word, sentence or document level. Fig 10 Opinion retrieval is a process of collecting reviews text from review websites. Information retrieval techniques such as web crawler can be applied to collect review text data from many sources and store them in database. This step involves retrieval of reviews, micro-blogs, comments etc of user. We should only consider the data which contain subjective data but not the objective data. Reviews are retrieved by query based information retrieval techniques. [4]

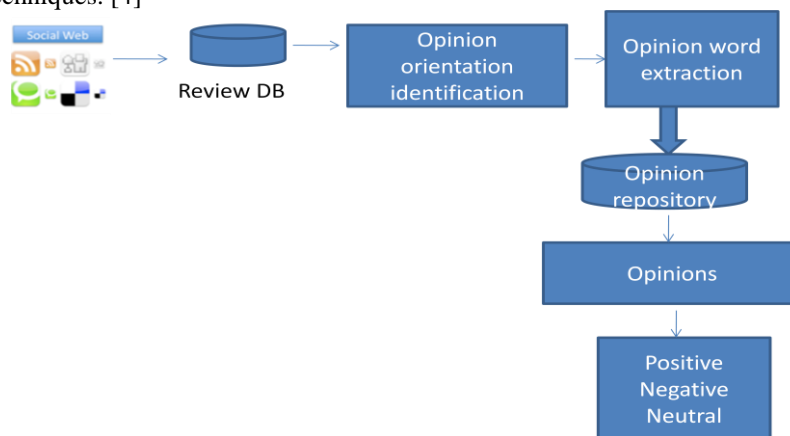


Fig 11 opinion mining process

## IX. CONCLUSION

In this project we investigated the learning capabilities of four machine learning methods for learning sentiment from students' textual feedback: Support Vector Machines (with three types of kernel). A dataset of 1036 instances of teaching-related feedback was used, which was labelled by 3 experts. We experimented with the use of unigrams as features and a range of standard pre-processing techniques. Our experiments indicate that two methods in particular, i.e. SVM with radial basis kernel and CNB, give very good results; therefore, they could be used for real-time feedback analysis. We also explored the use of the neutral class in the models and found that, in most cases, performance is better when the neutral class is not used. There are, however, arguments for using a neutral class from practical point of view, as it provides a more complete picture of a situation. Moreover, for the best performing method, i.e. SVM with radial basis kernel, the difference between using the neutral class and not using it, is 0.01 for accuracy, precision and recall. Consequently, one can argue that such a small loss is acceptable for having a more complete picture. Future work includes an analysis of more pre-processing techniques and their impact on model performance, as well as experimentation with other features, such as bigrams, trigrams and pos(part of speech)-tagging. In addition, we will test the models using more real-time collected data.



## REFERENCES

- [1] Sentiment analysis in Facebook and its application to e-learning Alvaro Ortigosa †, José M. Martín, Rosa M. Carro Department of Computer Science, Universidad Autónoma de Madrid, Francisco Tomás y Valiente 11, 28049 Madrid, Spain Computers in Human Behaviour 31 (2014) 527–541.
- [2] “Stumbl: Using Facebook to collect rich datasets for opportunistic networking research” May 2010 – October 201 M.Sc. Thesis Author: Georgios Nomikos vol-2 pp-38.
- [3] (2001). Adaptive hypermedia. User Modeling and User-Adapted7 Interaction, 11(1/2), 87– Brusilovsky, P., Schwarz, E., & Weber, G. (1996). ELM-ART: An intelligent tutoring system on World Wide Web. In Proc. of 3rd international conference on intelligent tutoring systems, ITS-96 (pp. 261–269). Montreal: Springer.
- [4] Survey on Opinion Mining and Summarization of User Reviews on Web, Vijay B. Raut et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1026-1030.
- [5] Sentiment Analysis on Big Data Machine Learning Approach span white paper.
- [6] Real Time Sentiment Analysis of Twitter Data Using Hadoop Sunil B. Mane et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3098 – 3100 3098
- [7] REST Representational State Transfer Michael Jakl mj@int-x.org 0226072 – 033-534 University of Technology Vienna
- [8] Scalable Sentiment Classification for Big Data Analysis Using Naïve Bayes Classifier Bingwei Liu\_Email:fbingwei.liu, dshen, gcheng@intfusion tech.comEmail:erik.blasch.1@u.af.milvol.3, 2013 (pp. 519–528).
- [9] Sentiment Analysis: Determining People's Emotions in Facebook N. AZMINA M. ZAMANI1,1Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA 40450 Shah Alam, Selengor MALAYSIA namz.ina@gmail.com, {zaleha, nasiroh} @tmsk.uitm.edu.my, mzabiden009@gmail.com.
- [10] Big Data and Sentiment Analysis using KNIME: Online Reviews vs. Social Media Ana Mihanović, Hrvoje Gabelica, Živko Krstić Poslovna inteligencija d.o.o., Zagreb, Croatia {ana.mihanovic, zivko.krstic}@inteligencija.com vol. 2, No. 1–2, 2008, pp 1-135.
- [11] Sentiment Analysis and Big Data Processing Mo Karimkhan, Jitendra B. Bhatia\ Computer Engineering, Insititute of technology, Nirma Universitykarimkhan\_it@yahoo.com.jitendra.bhatia@nirmauni.ac.in **IJCSC** Volume 5 • Number 1 March-Sep 2014 pp. 136-142 ISSN-0973-7391\_ 136