Text Mining in Analyzing the Presentation of Educational Trainers

S. Hari Ganesh, PhD.
Assistant Professor,
PG Research Department of
Computer Science,
Bishop Heber College,
Bharathidasan University,
Tiruchirapalli,620017,
TamilNadu,India.

ABSTRACT
This work deals with Text analysis that involves information retrieval through lexical analysis to learn word occurrence and distributions, pattern recognition, information extraction, data mining techniques and followed by visualization, and predictive analytics. The primary goal is to turn text into data for analysis, through application of natural language processing (NLP) and analytical tools. The problem taken for study is to evaluate a trainer through his lecture given in the class applying an innovative algorithm to perform the task.

Keywords
Data mining, Text mining, NLP

1. INTRODUCTION
In core the inspiration behind the expert area of content minin g, together with both text mining and data mining. known that the emergence of Natural Language Processing (NLP) during the 70's and 80's of the last century, text mining has been about in many forms and has evolved to a great extent in its techniques and its ambitions [7]. In the similar period, the most significant part of data mining received an optimistic impulse from artificial intelligence and neural networks applications [9]. In the nineties, content mining has given the foundations for better information recovery and search. More recently the attention has stirred towards semantic applications that take the significant context of information into account. In today's world text mining is a very important concept that used in Analysis and prediction. The work makes an evaluation of faculty performance using the concept of text and content mining. The voice of faculty is converted into text format and kept in another file. The text file which contains several keywords related to subject which faculty undertaking. Making correct analysis of text and context spoken by the faculty that correlates with the keyword in file, this paper can predict the teaching capability and presentation. The keywords of the subject are being stored in a file. The file consists of the text. The faculty's voice in the class room is been recorded and it been converted into text file that is been sent to server through the Bluetooth technology followed by the transcription services. The every word in the file contents are compared with the keywords that are stored in file, which has the list of key words of the subject taken, calculates the time that is being spent in every keyword. The study is being made related to Duration of time spent in each keyword are being recorded, it is repeated until all match with all the words spoken by the trainer compared with keywords in the file.

2. METHODOLOGY
In this paper it uses the concept of Bluetooth technology used to record the voice in the audio files which then converted into text file using transcription services. The key word file was created carefully using Experts in that domain which has to be incorporated with faculty's audio file. The analysis has been made as how many keywords spoken by the faculty for how many minutes that helps in evaluation of the faculty with the feedback of the faculty that had been already taken.

2.1 Trainer's Evaluation model-algorithm
Step 1: Start of lecture session.
Step 2: The Voice of the faculty recorded as audio file using WI_FI (blue tooth) technology and received in the nearest network server.
Step 3: The voice (Audio) contents are then changed to text format in to text file (Transcription services). Then the voice is transformed into text. The Semantic analysis was made and lexemes are generated.
Step 4: There is an Expert key file which contains 'Key' keywords opened.
Step 5: The "$Word_i$" where i=1, 2, 3, and 4…n word is taken that was counter checked with keyword file Key_i. If it matches count starts for the particular keyword 'i' using pattern matching algorithm like Brute force and Knuth Morris algorithms.
Else
The next keyword to be matched was tried, go to step 4.
Step 6: The count stops when the words of input file was exhausted.
Step 7: The time Duration far each keyword "Key_i" is "Timei" time which is spent for teaching was recorded.
Step 8: Analysis made with variable Time duration which is "$Time_i$" which is the total number in minutes which is used for computing performance "$Perform_i$" which will be compared with the student's feedback which is already been taken manually.
Step9: Analysis is made to predict the result of the trainer's performance.
Step 10: End of Trainer's Evaluation.

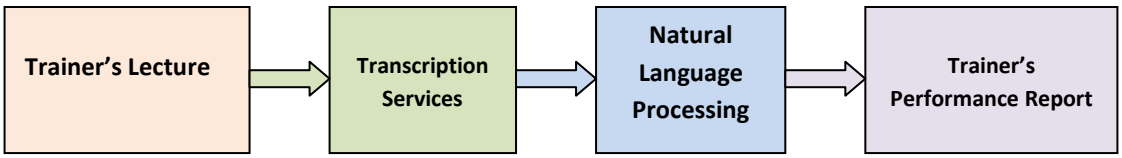

Fig 1: Architectural design of Trainer's Evaluation Model

2.2 Expertise of natural language processing

The process of document data is realized by means of grapheme, morphological [9], followed by syntax analysis and logical analysis and, semantic analysis, furthermore the initial texts are also indexed and remain in data base as documents [11][2].

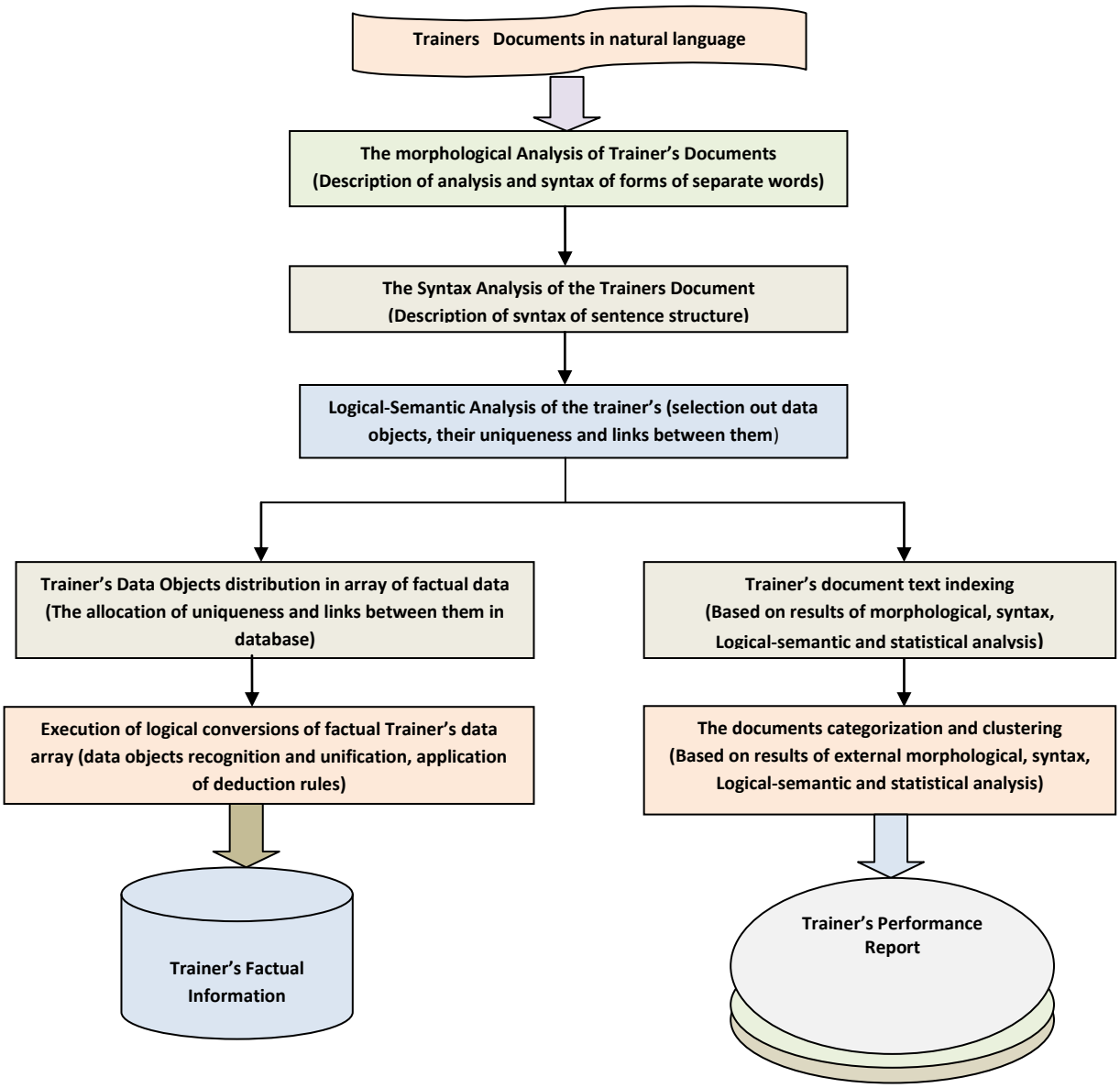

Fig 2: Technology of data processing in natural language.

2.3 First stage Grapheme analysis

The stage of grapheme analysis breakup into lexemes without morphology includes text fragments to separate recovery of special types like a word in quotes, a word, a punctuation symbol, a numerical block, an alphanumeric block and so on. The following distinctiveness are marked out for every fragments position, length and priority. Furthermore such features are represented in the register they are written in, are marked out for these examples. As consequence of this work is system of divide lexemes with number in text and without sub division of word [6]. If a word had several variations of parsing, then numerous fragments of system keep up a correspondence to it with one number as an example illustrating data processing sample data was taken. Eg. Web mining is a type of data mining used in customer relationship management and it takes huge amount of information.

The grapheme analysis for text is presented by the following way:

WORD1 (0, 1, 1, DATA, UpLw, 706)
WORD1 (1, 1, 1, MINING, Lw, 714)
WORD1 (2, 1, 1, IS, Lw, 722)
WORD1 (3, 1, 1, KNOWLEDGE, Lw, 730)
WORD1 (4, 1, 1, DISCOVERY, Lw, 738)
SPEC_SYMBOL (5, 0, 1, NEWLINE, 746)
NUM_BLOC (5, 1, 1, 03/24/2006, 753)
SPEC_SYMBOL (6, 0, 1, NEWLINE, 760)
WORD1 (6, 1, 1, WEB, UpLw1, 768)
WORD1 (7, 1, 1, MINING, Lw1, 776)
WORD1 (8, 1, 1, IS, Lw1, 784)
WORD1 (9, 1, 1, A, Lw1, 792)
WORD1 (10, 1, 1, TYPE, UpLw1, 800)
WORD1 (11, 1, 1, OF, Lw1, 808)
WORD1 (12, 1, 1, DATA, UpLw1, 816)
WORD1 (13, 1, 1, MINING, Lw1, 824)
WORD1 (14, 1, 1, USED, Lw1, 832)
WORD1 (15, 1, 1, IN, Lw1, 840)
WORD1 (16, 1, 1, CUSTOMER, Lw1, 848)
WORD1 (17, 1, 1, RELATION, Lw1, 856)
WORD1 (18, 1, 1, SHIP, Lw1, 864)
WORD1 (19, 1, 1, MANAGEMENT, Lw1, 872)
WORD1 (20, 1, 1, AND, Lw1, 880)
WORD1 (21, 1, 1, IT, Lw1, 898)
WORD1 (22, 1, 1, TAKES, Lw1, 910)
WORD1 (23, 1, 1, HUGE, Lw1, 918)
WORD1 (24, 1, 1, AMOUNT, UpLw1, 924)
WORD1 (25, 1, 1, OF, UpLw1, 933)
WORD1 (26, 1, 1, INFORMATION, UpLw1, 945)
PUNC_SYMBOL (27, 1, 1,., 987)

Note: Up1, Lw1, UpLw1 are registers

2.4 Second Stage Morphological analysis

During the stage of external morphological analysis all the words from the text was analyzed. A number in the character, an initial form, parts of speech, structure in which the word was met in the text and morphological characters were written then for each word of the text, as a consequence of morphological analysis was a system, detecting separate lexemes with their individual number in the text and without subdivision of abbreviation. If a word had more than a few variations in analysis, then it is results in a number of fragments of the system with single number.

START_FORM_MORF1 (0, 1, 1, DATA, , 768)
START_FORM_MORF1 (0, 1, 1, MINING, , 758)
START_FORM_MORF1 (0, 1, 1, WAS, ,788)
SUBJECT1 (1, 1, 1, KNOWLEDGE, 2780)

2.5 Syntax analysis

The syntax analysis stage of the text sentences was processed in sequence. Sentences constituted of order of words, symbols like punctuation mark and special series of symbols like different number formats, a collection of Latin lettering etc. afterward in this Stage It was used in resolve the fact and colors of relations between noticeable objects [4].

2.6 Logical-semantic analysis

On foundation of syntactical analysis the ultimate structures of texts are changed into a semantic network which nodes are represented by a large number of frequent terms, lexemes and set expressions. These network nodes are connected relatively between each other with dissimilar strength, which relies on frequency of occurrence of concepts in the sentences of the text [1][5]. This semantic network was used as a model of data domain in processing new unidentified documents.

- Analysis pertaining to domain objects
- Establishing connections between distinct objects

2.7 Selection out lexical definition

For selection out lexical definitions from lexemes preformed dictionaries are usually used. Structure of these dictionaries is a set of couples result oriented decoding, where result is an abbreviation or a distinct instance of any object. It is possible to have more than a few sources for an object dictionary and more than a few decoding for abbreviations.

Common dictionary structure looks like this:
Example: Abbreviations dictionaries:
<ITEM1>
<SRC1>abbreviation</SRC>
<DEST1>decoding_1</DEST1>
<DEST2>decoding_2</DEST2>
</ITEM1>
Example: Objects dictionaries
<ITEM2>
<SRC2>object_name_1</SRC>
<SRC2> object_name_2</SRC>
<DEST3>decoding</DEST3>
</ITEM2>
An example of an object dictionary:
<ITEM3>
<SRC3>MINING</SRC>
<SRC3>…</SRC>
<SRC>DATA</SRC3>
<SRC3>…</SRC3>
<SRC4>KNOWLEDGE</SRC4>
<DEST4>WORD</DEST4>
</ITEM3>

As a outcome of processing the lexeme, which was taken as an example, the lexical definitions as follows
WORD 1(12, 1, 1, DATA, 32019)
FIELD_STRUCTURE (24, 1, 1, MINING, 3946)
INO_NAME (24, 1, 1, OF, 3220)
These are the outcome of picking out lexical definitions

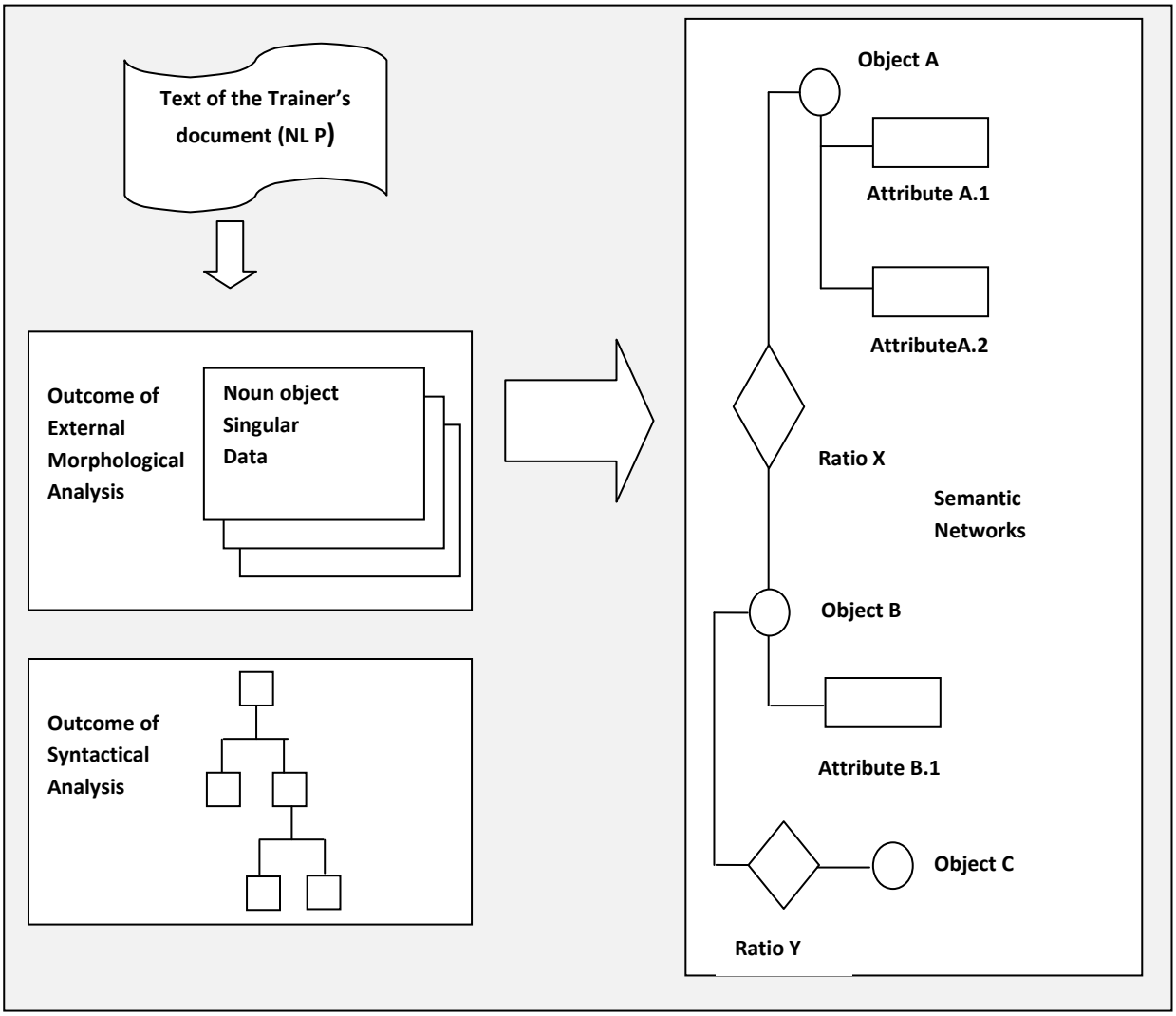

Fig 3: Depicts the process of morphological analysis

2.8 Identifying (marked) out domain objects

As parsing through domain objects, the ultimate set of objects is identified directly by the rules out of logical deduction. Even as analyzing objects good rules are progressed in sequence (good functions are called). Each function marks out essential lexemes by using its individual set of rules and after that it creates an original object on the support of these lexemes, saving essential data in the object. Every the processed lexemes was marked in a particular way to exclude a probability of reprocessing. This group of rules is called a first level set of rules. The general structure is as follows:

<RULE1 priority = "n">
<!—inquiry sub division -->
<QUERY1>OBJECT_NAME_1 (list of object parameters)</QUERY1>
.....
<!—blocks of OR QUERY -->
.....
<IGNORE1> OBJECT_NAME_N(list of object parameters) </IGNORE1>
.....
<QUERY2> OBJECT_NAME _N (list of object parameters) </QUERY2>
<!--functions division -->
<FUNCTION2> function_name_1 (function parameters) </FUNCTION2>
.....
<FUNCTION1> function name_N (function parameters) </FUNCTION1>
<!—action performed division -->
<CREATE1> OBJECT_NAME _1 (list_of_object_parameters) </CREATE1>
<INHERIT1> inquiry_number </INHERIT1>
<CREATE2> ATTRIBUTE (attribute_ name_1_1, meaning) </CREATE2>
.....
<CREATE3> ATTRIBUTE (attribute_ name_1_N, meaning) </CREATE3>
.....
<CREATE4> OBJECT_NAME _N (list_of_object_parameters) </CREATE4>
<CREATE5> ATTRIBUTE (attribute name _N_1, meaning) </CREATE5>
.....
<CREATE6> ATTRIBUTE (attribute name _N_N, meaning) </CREATE6>
<DESTROY1> INITIAL_FORM ($x,$y) </DESTROY1>
</RULE1>

2.9 The outcome of processing Text document information in the process of knowledge discovery

Process of detecting information in NLP, which was described above, was usually the first stage of text document information processing. The common outcome of first processing in the experimental system is an array of constrain factual data, offered as a semantic network, for identified marked of data objects from dissimilar sources, an identifying process was held out. It helps to tell related data objects which were got from dissimilar sources. When detecting the objects, two fundamental types of links that are chosen, links of resemblance and links of chance but at the same time there is a possibility of routine junction of concurrent objects is provided. Links of resemblance are usually processed by a system analyst (system analyst uses the expertise to locate if the data objects was concurrent and, if it is essential he decides to find their link by himself) before starting process of knowledge extraction. There was an option of correlate data objects that are recently placed in a factual base to the ones that are previously obtainable was considered to be an important aspect of detecting process. It allows solving troubles of monitor standard situation. When detection and link of current data objects are done, then array of real factual data was ready for knowledge discovery [10].

This procedure using different processes of real factual data:

Searching process: It uses Field search, dissimilar search, a full text search of similar documents Analytic process: It uses context dependent analysis, situational analysis, searching for communication series [3].

Process of modeling: simulation, prediction of situational development in long time.

These processes of data usually used in solving applications for supporting assumption, monitoring situation and analytic researches and so on.

3. RESULTS AND DISCUSSIONS

The Sample data are taken from real time class situations. The tables that show how analysis was made, Table 1 records the staff name with subject handled. The Table 2 displays keywords related to respective subject. The process involved that is shown in Table 3.

Table 1. Faculty Table (Sample Data)

Name of Faculty	Subject handled
James	Microprocessor
Davidson	Software Engineering
Karthikeyan	Networks
Divya	Mobile Computing
Shebba	Digital Fundamentals

Table 2. The Keywords Table (Sample Data)

Subject Domain	Keywords	Time (min.)
Microprocessor	Evaluation of Microprocessor	15
	Microprocessor Architecture	20
	Addressing Mode	10
	Instruction set	20
Software Engineering	Software Eng Definition	10
	Size Factors	15
	Quality and Productivity factors	20
	Preliminary Planning	10
	Walk through	10
Networks	Networks Definition	10
	LAN	10
	MAN	10
	WAN	10
	Topology definition	10
Mobile Computing	Mobile device definition	10
	Guided media	10
	Unguided media	10
	WI-FI	15
	Emulators	5
Digital Fundamentals	Number System	15
	Basic Gates	15
	Universal Gates	20
	Flip Flops	10
	Arithmetic Circuit	15

Table 3. Faculty presentation Table (Sample Data)

Name	Subject	Keywords	Time management Vs keywords
James	Micro-processor	Evaluation of Microprocessor Microprocessor Architecture Addressing Mode Instruction set	15-10 20-15 10-0 20-9
Davidson	Software Engineering	Software Eng Def. Size Factors Quality and Productivity factors Preliminary Planning Walk through	10-15 15-5 20-10 10-8 10-12
Karthikeyan	Networks	Networks Definition LAN MAN WAN Topology definition	10-7 10-10 10-5 10-12 10-20
Divya	Mobile Computing	Mobile device def. Guided media Unguided media WI-FI Emulators	10-10 10-9 10-12 15-13 5-5
Shebba	Digital Fundamentals	Number System Basic Gates Universal Gates Flip Flops Arithmetic Circuit	15-7 15-10 20-4 10-15 15-10

E=Time management * n
Performance= (E/n) *100
E=Earned points
N= Number of keywords

In Figure 4 Depicts staff performance related algorithm. Figure 5 represents the student's feedback which being conducted manually.

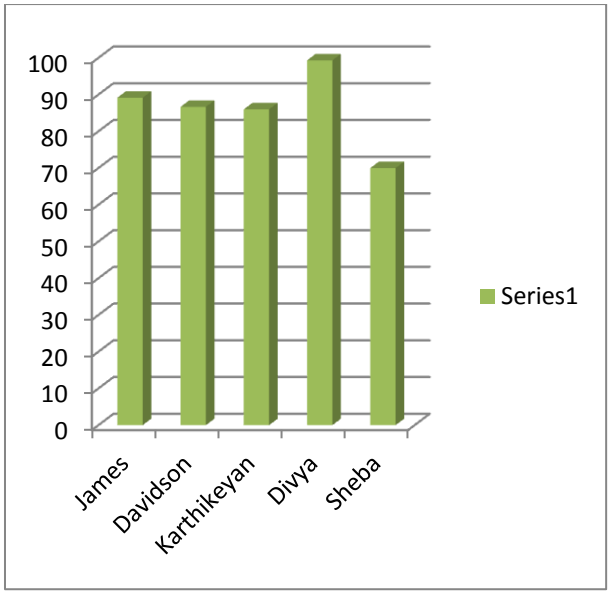

Fig 4: Depicts the staff performance by the innovative model

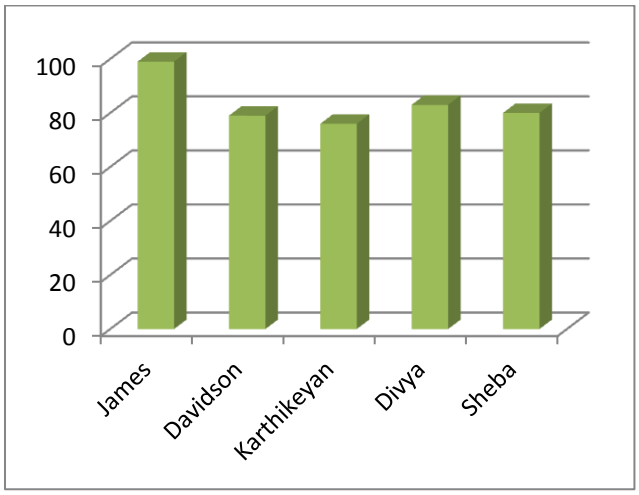

Fig.5: Depicts the student feedback taken manually

4. CONCLUSIONS

The Methodology adopted in this paper has provided opportunities to evaluate the performance of a trainer using the concept of KDD in teaching methodology. In evaluation model the voice of the trainer is being recorded and converted into text files and stored in a nearest server using the concept of Bluetooth technology and transcription services. The performance of the trainer is compared with manually taken feedback and the automated evaluation model, if there is standard deviation between the reports then there is an need for personal monitoring to be applied else the system is proper, so the management can take proper decision in assessing the performance of the trainer. In future more attributes can be added, like the trainers research attitude, body language and personality which make the KDD process more effective. In future it can be incorporated with Artificial Neural Network

5. REFERENCES

[1] David Traum, "Semantics and Pragmatics of questions

and answers for Dialogue Agents", proceedings of the International Workshop on Computational Semantics, pp. 380–394, January, 2003.

[2] Janda.M, "Grapheme based speech recognition", Proc. of the 18th Conference STUDENT EEICT 2012, Brno, CZ, VUT v Brně, pp. 441-445, 2012.

[3] Libossek M., Schiel F., "Syllable-based Text-to-Phoneme Conversion for German", Proc. ICSLP, Beijing, pp. 283-286, 2000.

[4] Mohamed Sherif, Axel-Cyrille Ngonga Ngomo, "Semantic Quran a Multilingual Resource for Natural-Language Processing", IOS Press, pp. 1-5,2003

[5] Navigli R, Velardi P, "Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation", IEEE Tran Pattern Anal. Intell, pp. 27(7), Mach, 2005.

[6] Raymond J. Mooney, Un Yong Nahm, "Text Mining with Information Extraction Multilingualism and Electronic Language Management", Proceedings of the 4th International MIDP Colloquium, Bloemfontein, South Africa, Daelemans, pp. 1-16, 2003.

[7] Ronan Collobert, Jason Weston, "A Unified Architecture for Natural Language Processing Deep Neural Networks with Multitask Learning", Proceedings of the 25 th International Conference on Machine Learning, Helsinki, Finland, pp. 1-8, 2008.

[8] Schwenk H., Gauvain, J., "Connectionist language modeling for large vocabulary continuous speech recognition", IEEE International Conference on Acoustics, Speech, and Signal Processing pp. 765–768, 2002.

[9] Siva kumar A. P., Dr. Premchand P., Dr. A. Govardhan, "Morphological Cross Reference method for English to Telugu Transliteration", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.2, No.4, pp 13-23, October 2011.

[10] Steyvers, Joshua, B. Tenenbaum, "The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth Mark", Cognitive Science vol.29, pp. 41–78, 2005.

[11] Tadesse Anberbir , Michael Gasser, "Grapheme-to-Phoneme Conversion for Amharic Text-to-Speech System", Conference on Human Language Technology for Development, Alexandria, Egypt, pp. 68-73, May 2011.