TEXT MINING OF JUDICIAL SYSTEM's CORPORA VIA CLAUSE ELEMENTS

Muhammad Ramzan Talib, Muhammad Kashif Hanif, Zubair Nabi, Muhammad Umer Sarwar, Nafees Ayub

Government College University, Faisalabad.
ramzan.talib@gcuf.edu.pk, mkashifhanif@gcuf.edu.pk, zubair.nabi@gcuf.edu.pk
mumersarwar@gcuf.edu.pk, nafees.ayub@gcuf.edu.pk
Pakistan.

Abstract: Legal professionals are interested to find, analyse, and reason about different previous cases drawn from a set of similar current cases. As a matter of fact, previous judgments those have been made on daily basis are part of the judiciary system to be used as previous reference by lawyers in their current cases. Although commercial sources of such legal information enables legal professionals looking case based on some specific keywords, however, the "casebase" And the search tools are inadequate. They are unable to automate a system that can understand, recognize and differentiate. This paper presents a methodology based on data and text mining techniques to support law practitioner and research scholars to trace desired information and identify all cases related to their relevant case.

Key words: Corpora, Text Mining, Data Mining, Natural Language Processing.

1. INTRODUCTION

In the latest world of technology the legal knowledge and information about judicial system is integrated and abundance in many formats. Large data sets that have information about judgment are either available in the form of judicial record books or available in the form of PDF or other simple text file. A legal case in general law terms is a dispute between divergent parties resolved by a judicial court. These resolved cases are then are part of judiciary data base duly maintained by the concerned authorities e.g. Caselaw Database, a US law firm JUSTIA [19]. From these datasets, it is quite difficult for researchers and legal professionals to get information of their own need with in very short period of time. Either they should read all the judgments one by one parsing line by line or they should study the

judgment books completely, which may lead to time consuming and depletion of very precious time.

In general law practice, different cases are referenced with respect to their relevant precedence and previous decisions rather than legislation. Legal professional like lawyers, justices, law students must find, analyses the different case factors and clauses and reason with and about cases drawn from a set of cases (previously judge), which is time consuming and hectic job which leads to bottleneck in text mining and information extraction [1, 2]. Although commercial providers of such legal information allow legal professionals and research students to search a case based on some specific keywords, however, the "case base" and the search tools are providing limited access to the data, non-extensible functionality, and have restricted access [9, 10].

This study is based on making classification to select a specific cluster as a test case to apply text mining techniques over judicial system corpora. These techniques will generate some of those results which will help in the analysis of case elements for judiciary system. This study will target to focus on developing a mechanism for fact findings in a scientific manner for facts uniformity, information extraction and text mining. This paper presents a methodology based on data and text mining techniques to support law practitioner and research scholars to trace desired information and identify all cases related to their relevant case.

In this paper, we apply natural language processing tools like, Parts of Speech (POS) Tagger, A Nearly New Information Extraction System (ANNIE) Gazetteer, ANNIE Orthomatcher, etc. JAVA Annotation Pattern Engine (JAPE) Transducer rules constructed to the textual elements in previous decided cases which are unstructured text, to produce annotated text for judiciary system for which text mining can be done and information can be extracted thus this technique can be used to solve the issue like bottleneck in text mining and information extraction [18].

In section 2 we discuss related work and literature review. In section 3, we elaborate our methodology, making of judicial system corpora of Supreme Court of Pakistan's previously judgement cases data using General Architecture of Text Engineering (GATE) [18]. In section 4 we conclude the paper, and discuss about future works.

2. RELATED WORK

In this section, we cover state of the art work on the legal case analysis and legal case base reasoning. There are two main branches about legal case based reasoning: one is knowledge representation and other is reasoning system which is constructed by manual analysis. Though, this division does not cover the area of knowledge bottleneck [1, 2]. There is another branch dealing with information extraction. This technique has issues about knowledge bottleneck using Natural

Language Processing (NLP) techniques. Previous some researchers worked over the ontology construction [3, 4], text summarization [5, 6], extraction of precedence link [7], factor analysis [8] and information extraction of case elements [9]. We focus on the clause clustering using judicial system corpora of Supreme Court of Pakistan's previously judged cases data. While working on the ontology, researchers believe that the focus should be defined or have information on the function of the law, leaving aside the high level of information, no-legal domain Information [10].

From unstructured data, text mining processes has been the subject of great attention of researchers over the past decade. In [11, 12] the authors present clarifications that deliver examples of structure for information extraction based on GATE. Various clustering methods [13, 14] can be used for this procedure. In most the cases, hierarchical clustering is used [15].

In [16] the author represents architecture about defining and validating a process of knowledge discovery. The author in [17] worked on the idea that multiple knowledge creation opportunities exist through the data mining process. The authors in [20] worked over the he lacks harmonisation of the rules regarding the protection of personal data in the police and criminal justice sector. The author in [21] worked over the Smart Metering Information System (SMIS) and presents some aspects of performance management of SMIS.

In the light of all above we focus our research work on text mining, information extraction and analysis using GATE tool [18]. As we initially based our work on the corpora selected from the data base of previously judged cases of Supreme Court of Pakistan and we start working with 99 cases drawn from the set of data available publically. For the feasibility of study, we give examples of some cases and provide result using judicial system corpora.

3. THE RESEARCH METHODOLY

In the methodology section, we used string processing with GATE software. GATE (General Architecture for Text Engineering) is open source application, software built in JAVA language that is being used in information extraction, text mining and text engineering. To complete our task, we have used different pattern evaluation resources like GATE document and grouped them in a specific corpus. After selecting documents from judicial record, GATE corpus initiates the judicial system data in form of combine corpora. Further, some more additional language processing tools such as, Document Reset PR-viouse (PR) for roll back, ANNIE English Tokeniser for lexical analysis, ANNIE Sentence Splitter for complex string processing, ANNIE POS Tagger for semantics, ANNIE Gazetteer for indexed terms, ANNIE Name Entity (NE) Transducer for pattern matching, ANNIE Orthomatcher for string grouping, has been implemented on the selected corpus. The processing rules and problem definitions are implemented in Java Annotation Pattern

Engine (JAPE) which has been constructed for the selected corpora. Figure 1 represents about the work flow diagram of the complete methodology using GATE [18].

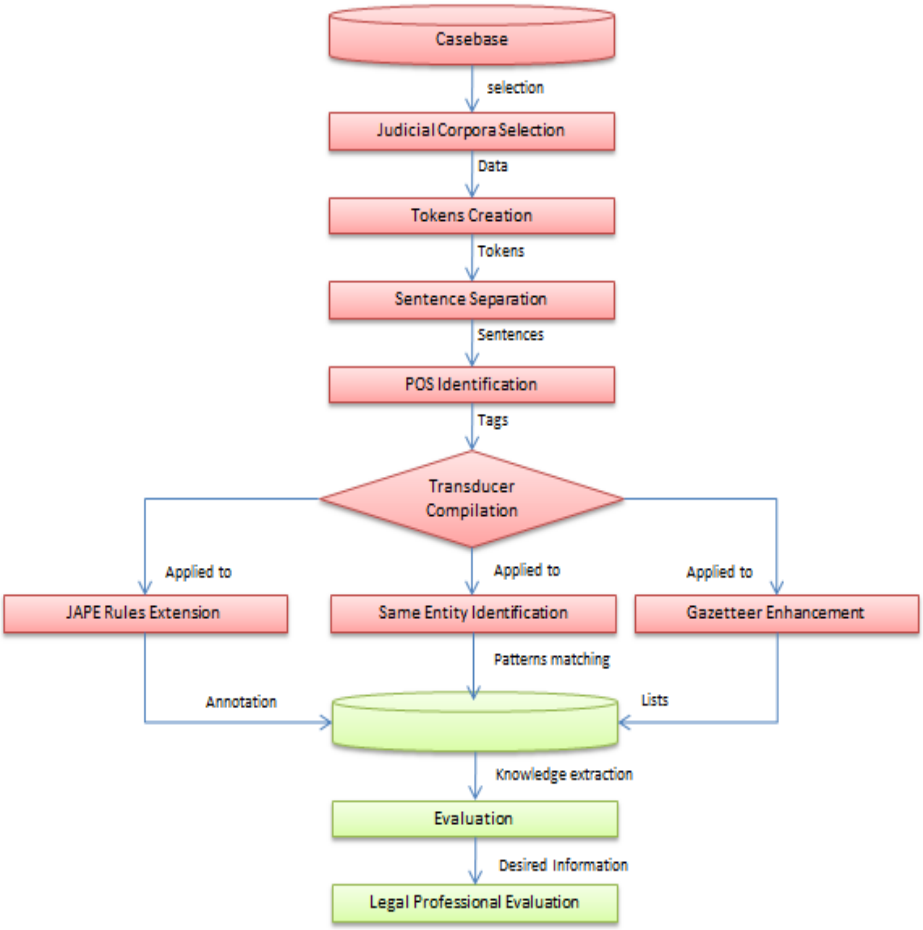

Fig. 1. Work Flow Diagram of the complete methodology using GATE

- **Judicial System Corpora.** To select the judicial system corpora from case base the set of 99 cases (previously judged) has been taken out.
- **Document Reset PRvious (PR).** Document Reset PR is used to reuse document for new rules and being used in the GATE to clear all the previous documents being loaded in the corpus area of the tool. We use the Documents Reset PR for more refinement of methodology phase.
- **ANNIE English Tokeniser**. ANNIE English Tokeniser used to manage case phrases for every character or space there is a token generated. Each

phrase being generated had different parameters like kind, length lower case, upper case, the word written in any form will be checked through this Phase.

- **ANNIE Sentence Splitter.** If we consider the parts of both type of language i.e. high level and low level, we share one commonality, this is the semantic of the language. The syntax form varies from language to language and type to type, but the semantics are of one type and there is nothing formal or informal in the semantics, to understand the semantic from the written text is possible through the processing of document. So, if we have a bulk of judicial data it is quite time consuming and difficult to process the textual data for a specific purpose. So, there is requirement of such methods which can perform the following identifications of the key terms. Interrelation of the judicial words, interrelation of the judicial sentences, the occurrence and positioning of the key terms.
- **ANNIE POS Tagger.** There are mainly two types of languages, the first language is formal language and the second language is informal language. Formal language is language which is used for the development purpose and informal language is human language like, English, French, German, Spanish, and Arabic. The most important thing about text engineering is the analysis of informal language with the help of formal language. The GATE provides us such facility to analyse informal language with the help of formal language.
- **ANNIE NE Transducer**. NE Transducer is used for judicial system as a pattern recognition small compilation part of judicial system processing. These are various patterns needed to be recognized as per the given case category. The pattern will define as problem definition and it can be recognized as the output or desired case result. The various components are related to the NE Transducer for efficient judicial pattern recognition. NE Transducer proved effective for legal clauses, as it can be enhanced for different legal clauses.
- **ANNIE Orthomatcher.** There are several connections in a document referring the same object, issue, person, place, event, entity etc. there are sometimes referred with key words or abbreviations. There are two ways to deal with them; one is the identification of them with same marker in alias ways. But there is also a way to join the defining new rules by extending the open source files. In our system, we can highlight the issue of same case in the bulk of data.
- **ANNIE Gazetteer.** ANNIE gazetteer is a list of lists that is already built in the open source application GATE, but for the specific purpose. Some built in files names are company, organization, day, date key, department,

hour, mountain, etc. These all built in file names are also saved in the "lists.def" file. To complete our task, we use the following path to reach at the gazetteer.

- **JAPE Transducer Rules.** Some JAPE transducer rules are already built in the application but to annotate some specific condition and information extraction from judiciary system we build our own rules.

Figures 2 and 3 present the rule for Judgement to annotate judgement of the cases and name of the Judges which appear in the different cases selected in the corpora.

```
Phase:firstpass
Input: Lookup
Options: control = brill
Rule: Judgement
Priority: 20
(
{Lookup.majorType == "Jud"}
): label
-->
:label.Judgement = {rule= "Judgement" }
```

Figure 2. Rule for Judgment Annotate the Judgment

```
Phase:firstpass
Input: Lookup
Options: control = brill
Rule: Present
Priority: 20
(
{Lookup.majorType == "Pre"}
): label
-->
:label.Present = {rule= "Present" }
```

Figure 3. Rule for Present Annotate the name of Judges

4. RESULTS

After completing methodology, in this section we discuss about the results of General Architecture for Text Engineering Tool (GATE). The below mentioned figures provides Information about the judgment orders made the honourable judges. In this picture the result is focus on the judgment. The highlighted portion in Fig. 4 clearly indicates the decision about the cases.

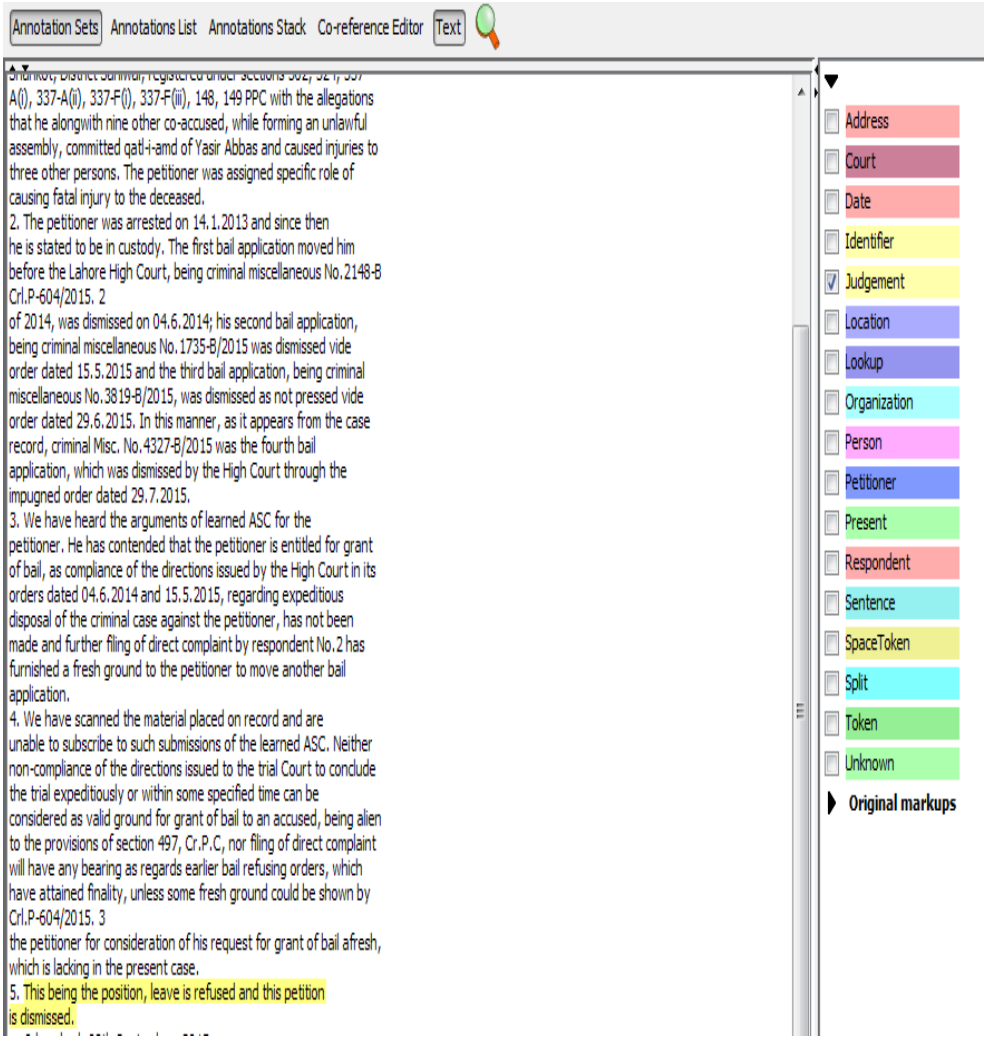

Fig. 4. Judgment information about the judgment orders

In Fig. 5 the result is focused on the highlighting the jurisdiction of the judgment and petitioner name and respondent etc. The light green colour of "present" tag clearly indicate the justices name in light green colour which the case was being present, furthermore court tag, petitioner tag and respondent tag clearly indicates the same colour information from the judicial system corpora.

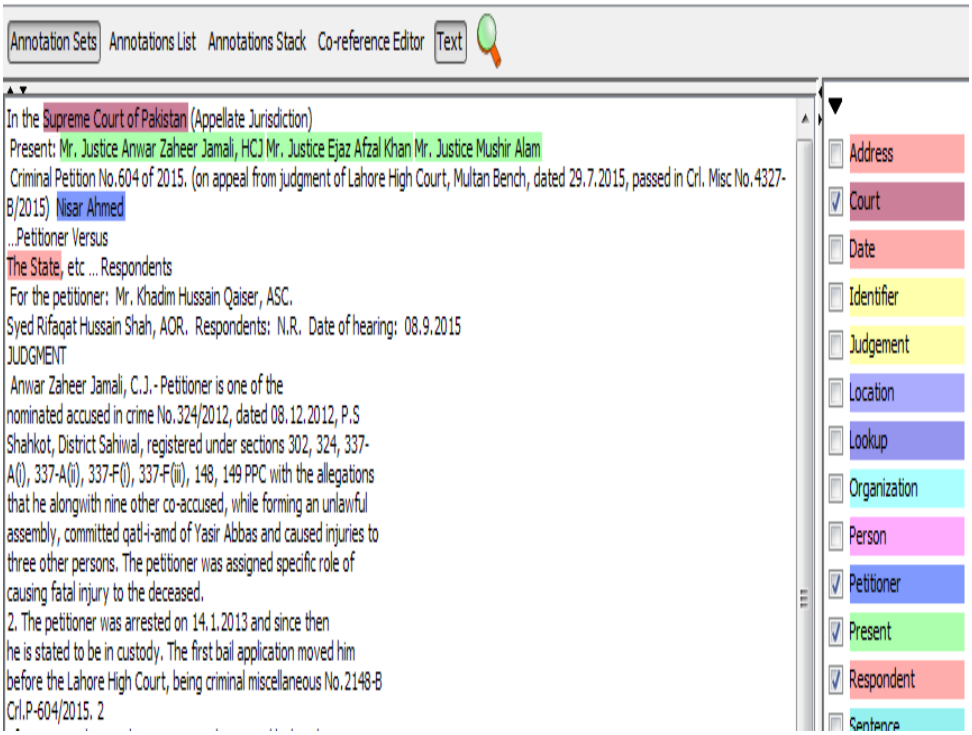

Fig. 5. Judgment information about court, petitioner, present and respondent.

Some Examples of Rules Representation.

Here are some examples of rule Representation built in JAVA language to get the required Information from the Judicial System Corpora. The below figures identify that how these rules can be constructed in the GATE.

```
Phase:firstpass
Input: Lookup
Options: control = brill
Rule: Petitioner
Priority: 20
(
{Lookup.majorType == "pet"}
): label
-->
:label.Petitioner = {rule= "Petitioner" }
```

Fig. 6. Example of Rule representation for Petitioner.

```
Phase:firstpass
Input: Lookup
Options: control = brill
Rule: Respondent
Priority: 20
(
{Lookup.majorType == "Res"}
): label
-->
:label.Respondent = {rule= "Respondent" }
```

Fig. 7. Construction of Rule Representation for Respondent

In the blew table some quantitative results have been shown with point of view that 99 number of cases has selected for the Judicial System Corpora publically available and some new tags has been generated in the GATE tool like judgement, Petitioner etc. By using all new tags for Information retrieval, we get some good results.

Table 1. For quantitative results with point of view # of doc.

Sr. No	**Corpora (Number of Cases)**	**New Tags**	**Information Retrieval**	**Success Rate in %**
1	99	Judgment	Yes	95%
2	99	Petitioner	yes	99%
3	99	Present	yes	98%
4	99	Respondent	yes	99%
5	99	Court Category	yes	98%

5. CONCLUSION

In this research article we discussed about the different annotation and how these annotations can further used for text mining using judicial System Corpora. However, this proposed study is feasible to overcome the bottlenecks in text mining and Information Extraction. This paper presents a methodology based on text mining techniques to support law practitioner and research scholars to trace out desired information and identify all cases related to their relevant case. A very small judicial system corpus applied to a very small ontology to get our desired results. Furthermore, this study can be extended to search engine optimization scheme for judicial system and legal text, not only this this research can further enhance to artificially intelligent judicial system.

Acknowledgments

We would like to present our thanks to president bar council, Faisalabad for recognition of our work and its need to the legislative matters. He helped us very much to arrange meetings with law firms and starting NLP for judicial system.

REFRENCES

[1] Ashley, K. *Modelling Legal Argument: Reasoning with Cases and Hypotheticals*. Bradford Books/MIT Press, Cambridge, MA, 1990.

[2] Rissland, E., Skalak, D., and Friedman, T. BankXX. Supporting legal arguments through heuristic retrieval. *Artificial Intelligence and Law*, 1996, 4 (1), pp. 1–71.

[3] Lame, G. Using NLP techniques to identify legal ontology components: Concepts and relations. *Artificial Intelligence and Law*, 2004, 12 (4), pp. 379–396.

[4] Peters, W. Text-based legal ontology enrichment. In P*roceedings of the workshop on Legal Ontologies and AI Techniques*, Barcelona, Spain, 2009.

[5] Moens, M.-F., Uyttendaele, C., and Dumortier, J. Abstracting of legal cases: the salom on experience. In *ICAIL '97: Proceedings of the 6th International Conference on Artificial Intelligence and Law*, New York, NY, USA, ACM 1997, pp.114-122.

[6] Hachey, B. and Grover, C. Extractive summarization of legal texts. *Artificial Intelligence and Law*, 2006, 14 (4), pp. 305–345.

[7] Jackson, P. Al-Kofahi, K., Tyrell, A. Vachher, A. Information extraction from case law and retrieval of prior cases. *AI*, November, 2003, 150 (1-2), pp.239–290.

[8] Ashley, K. and Brüninghaus, S. Automatically classifying case texts and predicting outcomes. *Artif. Intell. Law*, 2009, 17 (2), pp. 125–165.

[9] Wyner, A. and Peters, W. Towards annotating and extracting textual legal case factors. In *Proceedings of the Language Resources and Evaluation Conference Workshop on Semantic Processing of Legal Texts*, Malta, 2010.

[10] Wyner, A. and Hoekstra, R. A legal case OWL ontology with an instantiation of Popo v Hayashi. *The Knowledge Engineering Review*, 2010, 14. (2), pp.1-24.

[11] H. Cunningham et al., GATE: an Architecture for Development of Robust HLT Applications", In *ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2003, pp. 168-175.

[12] H. Cunningham et al. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics, *PLoS computational biology*, 2009.

[13] J. W. Raymond, C. J. Blankley, and P. Willett. Comparison of chemical clustering methods using graph- and fingerprint-based similarity measures, *Journal of Molecular Graphics and Modeling*, Mar. 2003, 21(5), pp. 421–433.

[14] J. MacCuish, C. Nicolaou, and N. E. MacCuish, Ties in Proximity and Clustering Compounds, *J. Chem. Inf. Computer. Sci.*, Jan. 2001, 41(1): 134–146.

[15] A. Böcker, S. Derksen, E. Schmidt, A. Teckentrup, and G. Schneider. A Hierarchical Clustering Approach for Large Compound Libraries, *J. Chem. Inf. Model Jul.*, 2005, 45(4), pp. 807–815.

[16] M. Castellano, N. Pastore, "A Flexible Mining Architecture for Providing New E-Knowledge Services" *IEEE 38th Hawaii International Conference on System Sciences*, 2005.

[17] G. Stephen, H. Meliha, "Knowledge Discovery through Visualising Using Virtual Reality" *IEEE 37th Hawaii International Conference on System Sciences*, 2004.

[18] The University of Sheffield. (1995-2016). (GATE) *General Architecture for Text Engineering*. Retrieved 20 March, 2016, from https://gate.ac.uk/

[19] Tim Stanley. (2016). *Justia*. Retrieved 31 March, 2016, from https://www.justia.com/

[20] Aced Félez, E. The Proposal of the European Commission for a Data Protection Directive in the Police and Criminal Justice Field. *International Journal on Information Technologies and Security*, ISSN 1313-8251, **2** (vol. 7), 2015, pp. 37-58. (http://ijits-bg.com)

[21] Atanasov, I. Modeling Aspects of Autonomous Smart Metering Information System. *International Journal on Information Technologies and Security*, ISSN 1313-8251, 1 (vol. 8), 2016, pp. 3-18. (http://ijits-bg.com)

Information about the authors:

Muhammad Ramzan Talib - PhD from Germany. I am also working as Associate Professor in Computer Science Department Government College University, Faisalabad. Areas of Research are Data Mining, Text Engineering.

Muhammad Kashif Hanif – PhD from Germany. I am also working as Assistant Professor at Computer Science Department Government College University, Faisalabad. Areas of Research are Big Data Analytics, Data Mining, Text Engineering.

Zubair Nabi – PhD Scholar, GC University, Faisalabad. I am also working as Lab Engineer in Software Engineering Department Government College University, Faisalabad. Areas of Research are computer Networks, Data Federation, Data Mining, Text Engineering.

Muhammad Umer Sarwar – PhD Scholar, GC University, Faisalabad. I am also working as Assistant Professor in Computer Science Department Government College University, Faisalabad. Areas of Research are Data Base Management System, Information Retrieval Data Federation, Data Mining, and Text Engineering.

Nafees Ayub – PhD Scholar, GC University, Faisalabad. I am also working as Lecturer in Computer Science Department Government College University, Faisalabad. Areas of Research are Data Base Management System, Information Retrieval Data Federation, Computer Networks and Security.

Manuscript received on 7 June 2017