AN ELABORATION OF TEXT CATEGORIZATION AND AUTOMATIC TEXT CLASSIFICATION THROUGH MATHEMATICAL AND GRAPHICAL MODELLING

Ahmed Faraz

Department of Computer & Software Engineering, Bahria University Karachi Campus,
13 National Stadium Road, Karachi -75260, Pakistan

ABSTRACT

As the time goes on and on, digitization of text has been increasing enormously and the need to organize, categorize and classify text has become indispensable. Disorganization and very little categorization and classification of text may result in slower response time of text or information retrieval. Therefore it is very important and essential to organize, categorize and classify texts and digitized documents according to definitions proposed by text mining experts and computer scientists. Work has been done on Text Mining, Text Categorization and Automatic Text Classification by computer and information scientists, but obviously a lot of space for novel research in this domain is available. In this paper we have proposed the mathematical notation and graphical models for Text Mining, Text Categorization and Automatic Text Classification to get in depth understanding of these techniques and concepts. Introduction and proposal of mathematical and graphical models for Text Mining, Text Categorization and Automatic Text Classification will shorten the response time of text and information retrieval. Also the performance of web search engines can be improved so much by employing these mathematical and graphical models.

KEYWORDS

Data Mining, Text Mining, Text Categorization, Text Classification, Automatic Text Classification, Text Spotting, Natural Language Processing, Knowledge Engineering, Knowledge Extraction, Information Storage/Retrieval

1. INTRODUCTION

In the last fifteen years, content-based document management system has obtained outstanding status in the field of Computer and Information Systems Engineering and Computer Science. There are two reasons for this popularity of content-based management system. The first one is that documents are available in digital form at a very large scale. The second one is that the human beings have natural desire to access them in a flexible way. Now we define Text Categorization which is also known as Text Classification or Topic Spotting. The significance of Text Categorization (TC) is that the most popular web search engines like Google, Yahoo, Alta vista, Web Searches, Bing and others use Text Categorization (TC) to search data and metadata through the employment of web crawlers and returns the optimal results. Also "Search Engine Optimization" is a newly emerging area of research in Computer Science which needs novel and advanced research in Text Categorization (TC). Consider an example of a room having a lot of things and accessories scattered in different directions. If one wants to search an item in this room he or she has to do a lot of efforts because of disorganization of items and human being's tendency to be confused by seeing a lot of things gathered together. If all the things are organized and placed on their appropriate locations, search will be easy and fast. The same example is

DOI : 10.5121/cseij.2015.5301

applied to search a text item from a pool of text. If the text is categorized and documents are classified among categories, search and retrieval of text will be fast and efficient.

2. TEXT CATEGORIZATION

We have been given a predefined set of natural language text then the method of labelling natural language texts with reference to thematic division is called Text Categorization (TC).There was an extensive work on Text Categorization in early 60s but this field was evolved gradually, and in early 90s it has gained prominent status and has become a major sub field of the Computer and Information Systems Engineering discipline. Obviously there is a role of increased power of software applications and the high availability of more powerful hardware in the emergence of Text Categorization (TC). There are now different applications of Text Categorization (TC) in many contexts. Some of the applications are Controlled Vocabulary Based Document Indexing, Document Filtering, Automated Meta Data Generation, Word Sense Disambiguation and Population of Hierarchical Catalogues of Web Resources. Generally speaking, Text Categorization (TC) is now being applied in multiple contexts covering any application requiring document organization, selective document dispatching and adaptive document dispatching. Text Categorization (TC) can be applied to the data which is in the form of natural language text. The natural language text is divided or categorized among subset of texts and labelled according to the theme which is the main idea or subject. Text Categorization is applied on online newspapers, online news channels, e-papers, web search engines because these web technologies incorporate search and retrieval of data in the form of text.

Let us consider an example of Text Categorization (TC).A newspaper web site say ABC wants to display three big news relating to nomination of a president of a country by the general assembly of that country, loss of financial assets of a firm and heavy rains in a specific region of country. Although huge amount of natural language text (in the form of news) is available for the web site but an efficient retrieval is required to access, retrieve and display the pertinent and specified text (news) on the main page of the news paper's web site. Obviously Text Categorization (TC) algorithm will be required to label the text according to the main subjects of text (i.e. theme).

2.1. Knowledge Engineering Approach

When we talk about real world applications of Text Categorization (TC) in the era from early 60s to late 80s, a lot of work had been done on Knowledge Engineering, which is an approach to Text Categorization (TC).The method adopted in Knowledge Engineering was that if someone wanted to classify documents under given categories, the experts knowledge was being encoded in the form of rules or a set of rules manually. In the 90s Knowledge Engineering approach lost its popularity so much and Machine Learning paradigm had gained more fame.

2.2. Machine Learning Paradigm Approach

The According to Machine Learning paradigm, there is a general inductive process which builds automatic text classifier itself through learning. The source of learning was originally a set of pre classified documents. From the given set of pre classified documents, characteristics of the categories of interest were learnt.

2.2.1. Advantages of Machine Learning Paradigm Approach

The advantages of the Machine Learning Paradigm Approach are that when we use it in Text Categorization then we do not need to get help from knowledge engineer or domain expert. When we compare Machine Learning Paradigm Approach with human experts, we gain accuracy. When

Machine Learning Paradigm Approach is used for the construction of the classifier or for its porting to a different set of categories, a huge savings is obtained in terms of human expert power. The reason is that there is no need of help or intervention from either the knowledge engineer or the domain expert.

2.3.A New Definition of Text Categorization

Finally current day Text Categorization is a discipline of Computer and Information Systems Engineering and Computer Science forming a sub set through intersection of two sets of Machine Learning and Information Storage/Retrieval (ISR).There is a sharing of a number of characteristics between Text Categorization (TC) and other tasks such as knowledge or information extraction from texts and text mining.

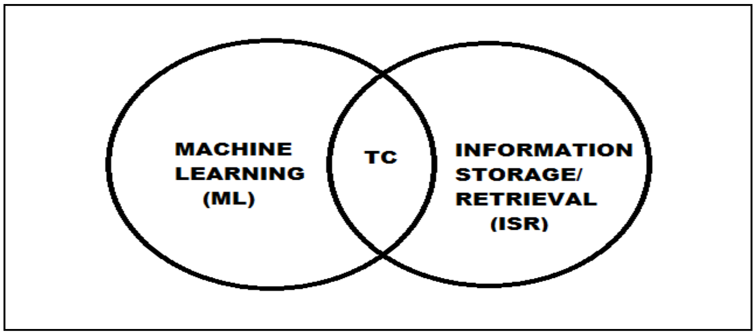

Figure 1: Set Notation of Text Categorization

Text Categorization (TC) is a newly emerging area of Computer and Information Systems Engineering and Computer Science and its terminology and notation is still evolving. It is very difficult for us to identify the borders of Machine Learning (ML) and Information Storage/Retrieval because most of the things and concepts are common to both disciplines.

2.3. Some Definitions related to Text Categorization

The following terms should be clearly understood by the readers to get thorough understanding of Text Categorization (TC).

2.4.1. Categories

The term ‘Categories’ is defined as symbolic labels and no additional knowledge of their meaning is available. The term ‘additional knowledge’ means knowledge of procedural or declarative nature.

2.4.2. Exogenous Knowledge

An external source provides data for the purpose of its classification .This data is called Exogenous Knowledge.

2.4.3. Endogenous Knowledge

The term ‘Endogenous Knowledge’ is defined as the knowledge acquired from the documents. It is assumed that no Exogenous Knowledge is available and only Endogenous Knowledge is available for Text Categorization (TC).

2.4.4. Metadata

The term ‘Metadata’ is defined as data about data. It is covered under the heading of Exogenous Knowledge. The example of Metadata includes a research paper available for download and reading on Google Scholar. This research paper has publication source, publication date, document type. This is apparent that the search engines only provide Exogenous Knowledge to the users or the search engines are the only source of Exogenous Knowledge.

3. TEXT MINING

The Text Mining can be defined in three steps theoretically. These three steps are implemented by Text Mining scientist as follows:

(a) Analysis of large quantities of text
(b) Detection of usage patterns from text
(c) Extraction of useful and correct information from detected usage patterns.

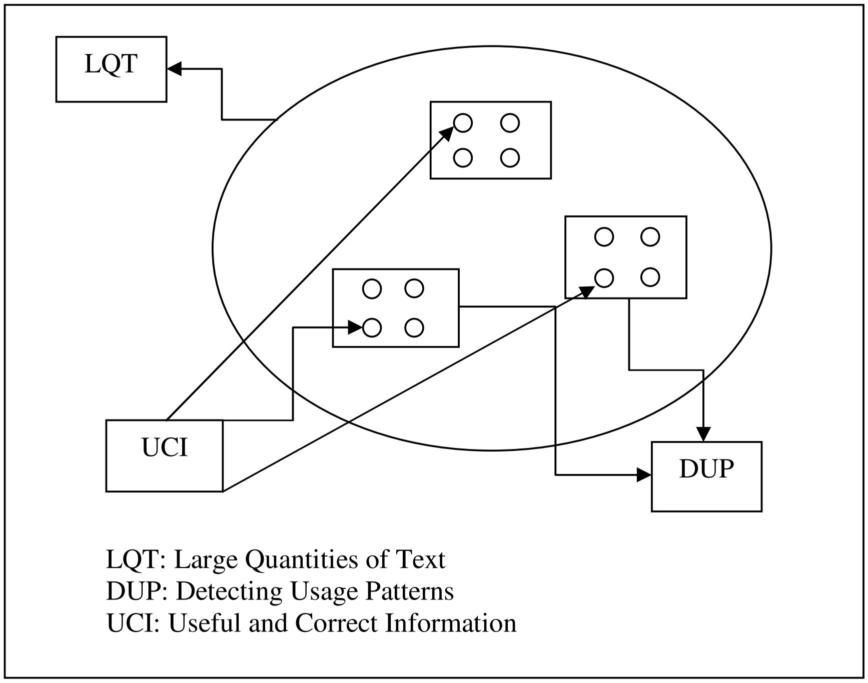
LQT
UCI
DUP
LQT: Large Quantities of Text
DUP: Detecting Usage Patterns
UCI: Useful and Correct Information

Figure 2: Text Mining Model

According to this definition, Text Categorization (TC) is an instance of Text Mining. Yet Text Categorization (TC) is an immature field of Text Mining and there is no systematic treatment of the subject. The field of Text Categorization (TC) is broad enough and requires a lot of work in the future and it does not contain a good collection of text books and journals. Two journals are dedicated for Text Categorization (TC).These journals are Joachims and Sebastiani (2002) [1,3], Lewis and Hayes (1994) [4,5].there is a term in the field or discipline of Text Mining which is “Automatic Text Classification” (ATC).

3.1. Automatic Text Classification (ATC)

The readers should have very clear concept in their minds that there is a difference between Automatic Text Classification (ATC) and Text Categorization (TC).We have proposed the new definitions of Automatic Text Classification (ATC) here which are different from the definitions from the literature. Also we have introduced novel mathematical notation and models for Automatic Text Classification (ATC).These novel definitions, mathematical notation and models are discussed below:

3.1.1.Definition (i)

Automatic Text Classification (ATC) can be defined as automatic assignment of documents to a predefined set of categories.

3.1.2.Definition (ii)

Automatic Text Classification (ATC) can be defined as automatic identification of such a set of categories “definition by Borko and Bernick(1963)”[6].

3.1.3.Definition (iii)

Automatic Text Classification (ATC) can be defined as automatic identification of such a set of categories and the grouping of documents under them “definition by Merkl(1998)”[7]. This work to achieve is called Text Clustering.

3.1.4.Definition (iv)

Automatic Text Classification (ATC) can be defined as any activity of placing text items into groups “definition by Manning and Schutze(1999)”[8]. This work includes both Text Categorization (TC) and Text Clustering as particular instances of Automatic Text Classification (ATC).

3.2. Mathematical Notation of Automatic Text Classification

Assume that the following mathematical notation is used to denote the concept of Automatic Text Classification (ATC):

$\{Ci\}$:A predefined set of categories (i^{th}category)
$\{Di\}$:Documents
$\{\leftarrow\}$:Assignment operator
$\{LQT\}$:Large quantities of text
$\{=\}$: contains
$\{T_i\}$: Text Items (i_{th} Text Item)
$\{G_i\}$: Groups (i_{th}Group)

3.2.1.Mathematical Notation of Definition (i)

Generally Automatic Text Classification (ATC) can be defined and represented mathematically as follows:

$$\{C_i\} \leftarrow \{D_i\} \tag{1}$$

When we consider specifically Automatic Text Classification (ATC), we have the following mathematical notation:

$\{C_1\} \leftarrow \{D_1\}$
$\{C_2\} \leftarrow \{D_2\}$

$\{C_3\} \leftarrow \{D_3\}$
$\dots\dots\dots\dots\dots$
$\{C_n\} \leftarrow \{D_n\}$

Where $i = 1 \dots\dots\dots n$ .Here$\{C_1\}$, $\{C_2\}$, $\{C_3\}$,………..….,$\{C_n\}$ represents first 1_{st}category, second2_{nd} category, third 3_{rd}category ,so on and n_{th}category respectively. And $\{D_1\}$,$\{D_2\}$, $\{D_3\}$,………….,$\{D_n\}$ represents 1_{st} document, 2_{nd} document,3_{rd} document so on and n_{th}category respectively.

By the application of definition (i) of Automatic Text Classification (ATC), Large Quantities of Text (LQT) changes the form from figure 3 to figure 4.

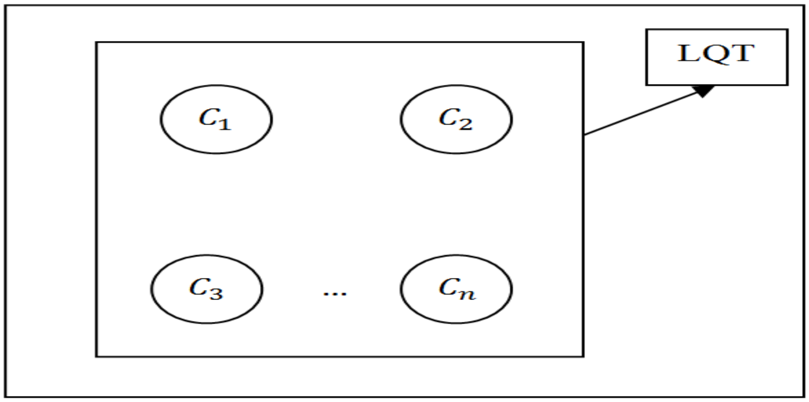

Figure 3: LQT before application of definition (i)

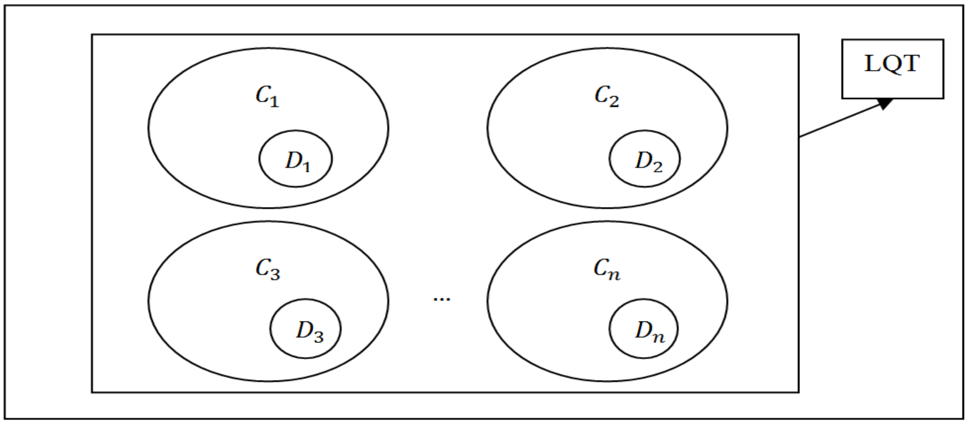

Figure 4: LQT after application of definition (i)

3.2.2. Mathematical Notation of Definition (ii)

Automatic Text Classification (ATC) can be defined and represented mathematically as follows:

$$\{LQT\} = \{C_1, C_2, C_3, \dots\dots\dots, C_n\} \quad (2)$$

Where $\{LQT\}$ represents Large Quantities of Text and $\{C_1, C_2, C_3 \ldots\ldots\ldots C_n\}$ represents a set of categories of documents. The algorithm for Automatic Text Classification (ATC) should be able to identify i_{th} category C_i from$\{LQT\}$.

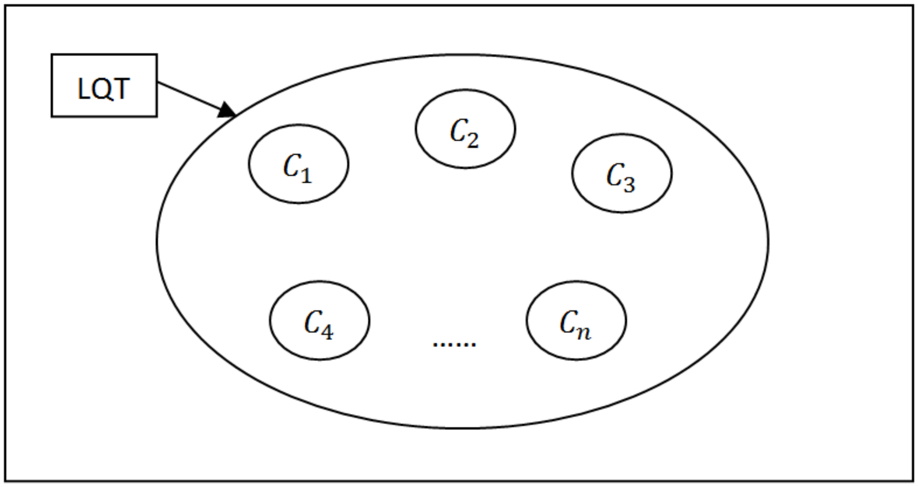

Figure 5: Automatic Text Classification Model (Definition ii)

Automatic identification of a set of categories is modelled as follows:

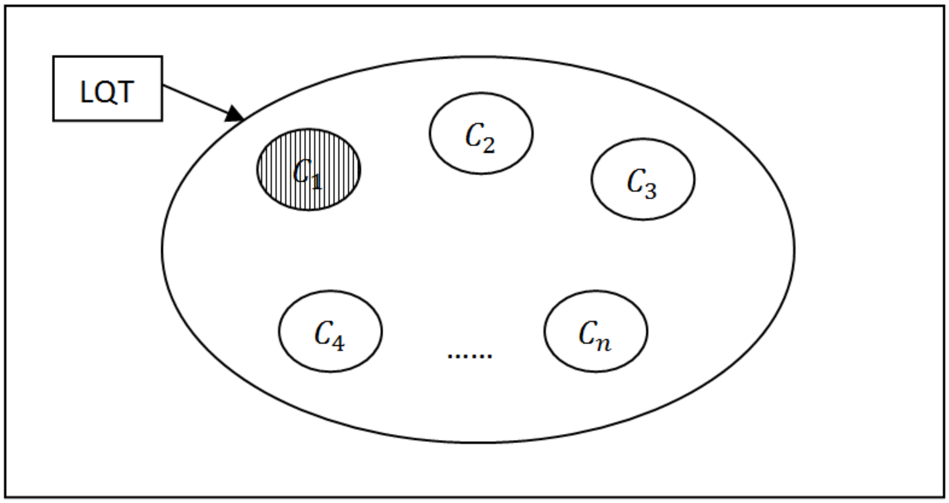

Figure 6: C_1 is identified among Automatic Text Classification (ATC) Model (Definition ii)

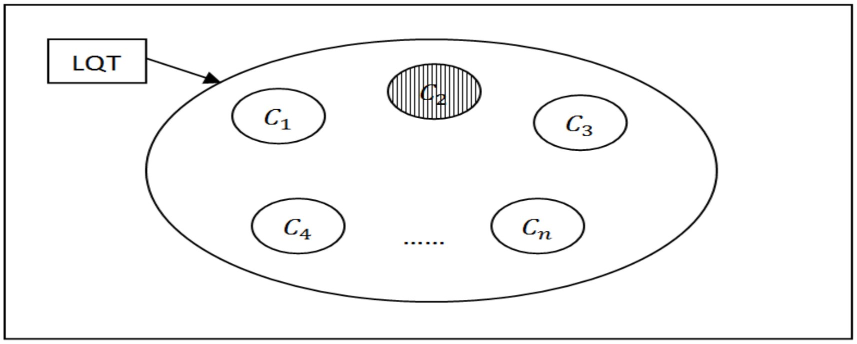

Figure 7: C_2 is identified among Automatic Text Classification (ATC) Model (Definition ii)

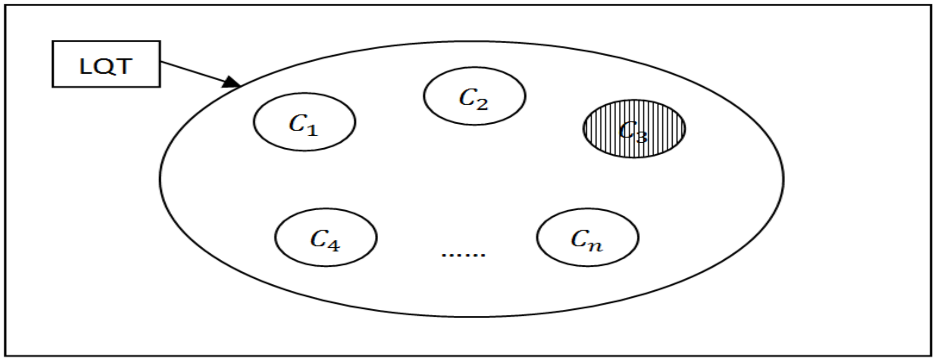

Figure 8: C_3 is identified among Automatic Text Classification (ATC) Model (Definition ii)

3.2.3. Mathematical Notation of Definition (iii)

Automatic Text Classification (ATC) can be defined and represented mathematically as follows:

$\{LQT\} = \{C_1, C_2, C_3, \dots \dots \dots, C_n\}$

Here a set of categories $C_1, C_2, C_3, \dots\dots\dots, C_n$ are automatically identified from Large Quantities of Text (LQT) and documents are grouped under them.

Category C_1 contains a group of documents A, B, C, D .

$C_1 = \{A, B, C, D\}$

Category C_2 contains a group of documents G, H, I, J .

$C_2 = \{G, H, I, J\}$

Category C_3 contains a group of documentsM, N, O, P.

$C_3 = \{M, N, O, P\}$

Category C_n contains a group of documentsW, X, Y, Z.

$C_n = \{W, X, Y, Z\}$

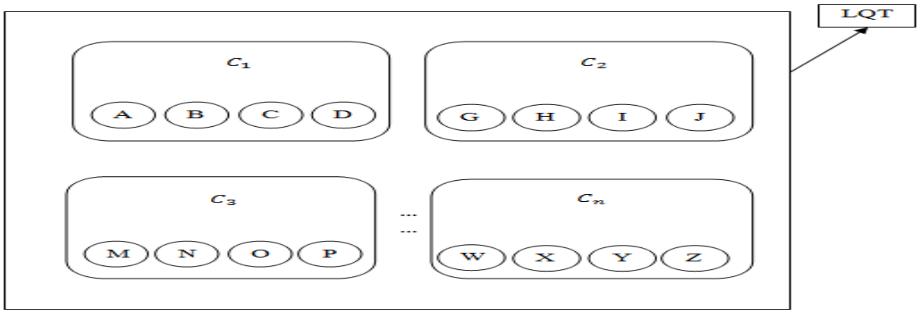

Figure 9: Automatic Text Classification Model (Definition iii)

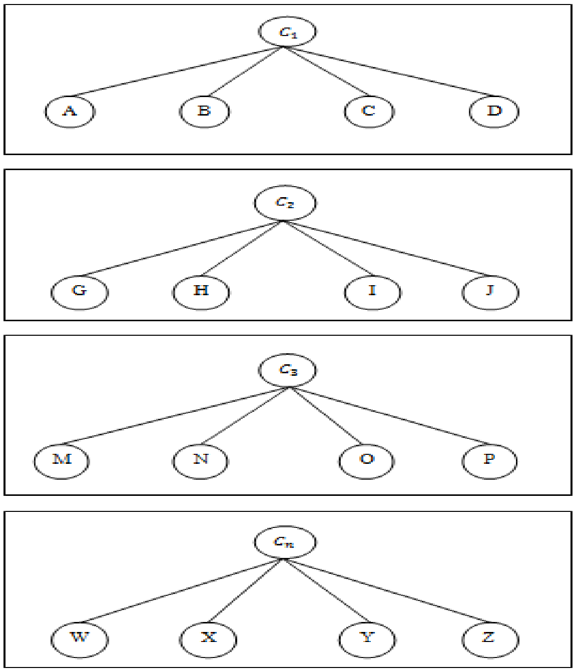

Figure 10: Category Identification and Document Grouping (Definition iii)

3.2.4. Mathematical Notation of Definition (iv)

Assume that text items are randomly assigned to groups. Generally Automatic Text Classification (ATC) can be defined and represented mathematically as follows:

$\{G_1\} \leftarrow \{T_1 , T_2\}$
$\{G_2\} \leftarrow \{T_3, T_4\}$
$\{G_3\} \leftarrow \{T_5, T_6\}$
$\{G_4\} \leftarrow \{T_7, T_8\}$
$\dots \dots \dots \dots \dots \dots \dots$
$\{G_n\} \langle \ \{T_{n-1}, T_n\}$

We have assumed that two text items are assigned to a group in our mathematical notation for the sake of simplicity, but more than two text items can be assigned to a single group. Also the point

to be noted here is that the mathematical notation of Automatic Text Classification (ATC) definition (iv) presented here is shallow. The reason is that definition (iv) includes the implementation of definition (iii), therefore it should be kept in mind that the implementation of mathematical notation of definition (iv) will require the implementation of mathematical definition (iii). For the purpose of placement of text items into groups, a group of documents should be identified and selected through Artificial Intelligence based technique. This problem is not addressed here and it would be the future research direction in the field of Text Mining.
Also a second question is raised which states that which text items are need to be placed on a given group. The selection of text items and identification of pertinent group remains a new direction of research in the field of Text Mining. Definition (iv) needs more work.

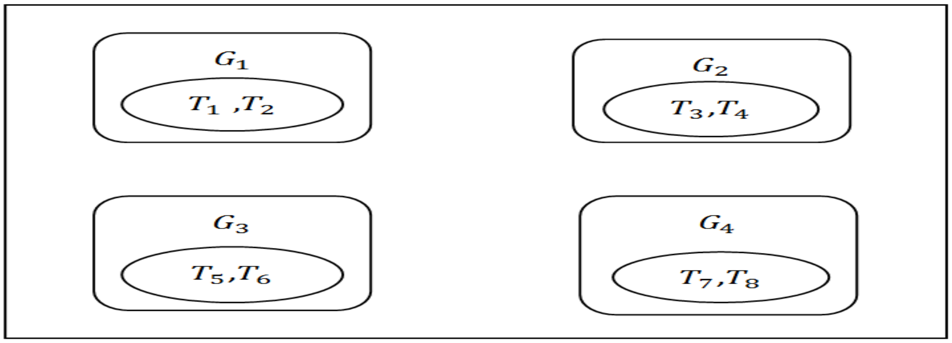

Figure 9: Automatic Text Classification Model (Definition iv)

4. FUTURE RESEARCH WORK

It Consider an example of a news published in the newspaper daily ‘Dawn’, “...”.This news may be defined in the category of sports or politics or in both or in neither of the category. It is the responsibility of human expert to identify the category for publishing news under which it should be covered .Automatic identification of text categories without involvement of human expert is the problem for future research work.

Secondly, there are two types of knowledge associated with a document defined under a specific category. The first one is ‘Exogenous’ and the second one is ‘Endogenous’. When we do Text Categorization (TC) and Automatic Text Classification (ATC), the main reliance is on Endogenous knowledge. The second research problem identified during our research work is the role and effects of exogenous knowledge in Text Categorization (TC) and Automatic Text Classification (ATC).

5. CONCLUSIONS

In this paper we have introduced the mathematical notations and graphical representations of definitions of Automatic Text Classification (ATC). Also we have developed Text Mining Model. This work will help to facilitate the design and development of algorithms for Text Categorization (TC) and Automatic Text Classification (ATC) which would improve the performance of Text Mining based softwares.

It can be deduced from the mathematical notation and diagrammatic representation of Automatic Text Classification (ATC) that the definition by Borko and Bernick (1963)[6] is extending the first definition, definition by Merkl (1998)[7] is extending the definition by Borko and Bernick

(1963)[6] and definition by Manning and Schutze(1999)[8] is the union of definition by Merkl (1998)[7] and definition by Borko and Bernick (1963)[6].

ACKNOWLEDGEMENTS

I would like to thank the Director General of Bahria University Karachi campus, Vice Admiral Khalid Amin HI(M) (Retired)and the Director of Bahria University Karachi Campus, Captain Mohsin H. Malik TI(M) PN for motivating me to be involved in research work in my field of interest. They have always persuaded the faculty to do research work for the sake of serving the humanity, science and engineering.

REFERENCES

[1] F. Sebastiani, "Machine learning in automated text categorization," ACM computing surveys (CSUR), vol. 34, pp. 1-47, 2002.
[2] F. Sebastiani, "Text Categorization," ed, 2005.
[3] T. Joachims and F. Sebastiani, "Guest editors' introduction to the special issue on automated text categorization," Journal of Intelligent Information Systems, vol. 18, pp. 103-105, 2002.
[4] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," The Journal of Machine Learning Research, vol. 5, pp. 361-397, 2004.
[5] D. D. Lewis and M. Ringuette, "A comparison of two learning algorithms for text categorization," in Third annual symposium on document analysis and information retrieval, 1994, pp. 81-93.
[6] H. Borko and M. Bernick, "Automatic document classification part II. Additional experiments," Journal of the ACM (JACM), vol. 11, pp. 138-151, 1964.
[7] D. Merkl, "Text classification with self-organizing maps: Some lessons learned," Neurocomputing, vol. 21, pp. 61-77, 1998.
[8] C. D. Manning and H. Schütze, Foundations of statistical natural language processing: MIT press, 1999.
[9] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to information retrieval vol. 1: Cambridge university press Cambridge, 2008.
[10] T. Joachims, Text categorization with support vector machines: Learning with many relevant features: Springer, 1998.
[11] T. Joachims, Learning to classify text using support vector machines: Methods, theory and algorithms: Kluwer Academic Publishers, 2002.

Author

Mr. Ahmed Faraz holds Bachelor of Engineering in Computer Systems and Masters of Engineering in Computer Systems from N.E.D University of Engineering and Technology, Karachi Pakistan .He has taught various core courses of computer science and engineering at undergraduate and postgraduate level at Sir Syed University, N.E.D University and Bahria University Karachi for more than ten years. His research interests include AI, Data Mining, Parallel Processing, CAO, Statistical Learning.