A COMPARISON OF TEXT CATEGORIZATION METHODS

Ahmed Faraz

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

ABSTRACT

In this paper firstly I have compared Single Label Text Categorization with Multi Label Text Categorization in detail then I have compared Document Pivoted Categorization with Category Pivoted Categorization in detail. For this purpose I have given the general definition of Text Categorization with its mathematical notation for the purpose of its frugality and cost effectiveness. Then with the help of mathematical notation and set theory ,I have converted the general definitions of Single Label Text Categorization and Multi Label Text Categorization into their respective mathematical representation .Then I discussed Binary Text Categorization as a special case of Single Label Text Categorization. After comparison of Single Label Text Categorization with Multi Label Text Categorization, I found that Single Label Text Categorization or Binary Text Categorization is more general than Multi Label Text Categorization. Thereafter I discussed an algorithm for transformation of Multi Label Classification into Binary Classification and explained the conditions of transformation of Multi Label Classification into Binary Classification. In the second step I compared Document Pivoted Categorization with Category Pivoted Categorization in detail. After comparison we found that Category Pivoted Categorization is more typical and complex than Document Pivoted Categorization. The Category Pivoted Categorization becomes more complicated when new category is added to predefined set of categories and the recurrent classification of documents takes place. Finally I compared Hard Categorization with Ranking Categorization. After comparing them I found that Hard Categorization incorporates 'Hard Decisions' about the relevance or belonging of a document to a category. This hard decision is either completely true or completely false. Whereas the Ranking Categorization creates a belonging of a document to a category according to the estimated appropriateness to the document. The final Ranked List is developed in the Ranking Categorization which is used by the human expert for final decision of Text Categorization.

KEYWORDS

Text Mining, Text Categorization, Automatic Text Classification, Single Label Classification, Multi Label Classification, Binary Classification, Document Pivoted Categorization, Category Pivoted Categorization, Hard Categorization, Ranking Categorization, Ranked List, Semi Automated Classification

1. INTRODUCTION

When we discuss Text Categorization, we elaborate it with the general context and we do not consider Text Categorization with reference to application or development point of view. Theoretically we define or give general definitions of Text Categorization because of the fact that it is more frugal and cost effective and it may be applied to any application in a specific domain. It is apparent that we do not have any need to verify these suppositions in a particular environment in which Text Categorization is implemented. It may be possible that when Text Categorization is implemented in a particular domain, any other source of information may be required to be available at the time of implementation of Text Categorization as component.

DOI: 10.5121/ijnlc.2016.5103

The essential condition for Text Categorization is that at the first stage, the source of text or the document on which Text Categorization is applied should be available. The availability of the document at the first stage is indispensable. Let us suppose that Text Categorization is applied on the text of a newspaper, and then the text of newspaper should be available at the first time. Similarly if Text Categorization is applied on the text of an e-book, then the text of e-book should be available at the first time. The second important point to discuss is that Text Categorization is purely a field of Text Mining and the Text Mining is a sub-field of Data Mining. The Text Categorization techniques and methods as discussed mainly in [3], [5], [9], [10] and other referred papers are applied on pure text and it does not apply on other types of data like audio, video, MPEG, MP3, audio streams from a source, video streams from a source etc. In order to compare different Text Categorization techniques, we define Text Categorization first of all. Text Categorization is the task of assigning a Boolean value to each pair [3].

$$(d_j, c_i) \in D \times C$$

Where D is domain of documents and C is a set of predefined categories

$$C = \{c_1, c_2, c_3, \dots \dots \dots \dots \dots \dots, c_{|C|}\}$$

2. Single Label Versus Multi Label Text Categorization

When Text Categorization is implemented different constraints may be applied on the Text Categorization task [2]. I have given the general definition of Text Categorization in the above paragraph. Consider general definition of Text Categorization for a particular case where $"k"$ is a given integer. The term “exactly k” means either $\leq k \;\; or \;\; \geq k$ is considered. From the general definition of Text Categorization we know that:

$$(d_j, c_i) \in D \times C$$

This implies that two memberships of documents and categories are derived:

$$d_j \in D$$

$$c_i \in C$$

Consider the first case in which exactly k elements of C will be assigned to each $d_j \in D$.
This means that either $\{c_1, c_2, c_3, \dots \dots \dots \dots \dots \dots, c_k\}$will be assigned to each $d_j \in D$ or $\{c_k, c_{k+1}, c_{k+2}, \dots \dots \dots \dots \dots \dots\}$ will be assigned to each $d_j \in D$.

The case in which exactly one $(\leq 1 \;\; or \;\; \geq 1)$category must be assigned to each $d_j \in D$is called Single Label Text Categorization .The Single Label Text Categorization is also called Non Overlapping Categories case.

The case in which any number of categories from 0 to |C| may be assigned to the same $d_j \in D$ is called Multi Label Text Categorization as defined and discussed in [8] .The Multi Label Text Categorization is also called Overlapping Categories Case as discussed in [1] and [8].

2.1. Binary Text Categorization

We define Binary Text Categorization as a special case of Single Label Text Categorization as discussed in [11].In Binary Text Categorization each $d_j \in D$ must be assigned either to the category c_i or its complement $\overline{c_i}$.

We can represent Binary Text Categorization mathematically as follows:

$$\forall\{d_j \in D\} \rightarrow c_i$$

$$\forall\{d_j \in D\} \rightarrow \overline{c_i}$$

2.2. First Comparison Result

When we compare Single Label Text Categorization with Multi Label Text Categorization we see that Single Label Text Categorization is more general than Multi Label Text Categorization theoretically. Since Binary Text Categorization is a special case of Single Label Text Categorization as illustrated in [1] , [8] and [11].We can state that Binary Text Categorization is more general than the Multi Label Text Categorization theoretically.

2.3. Use of Algorithm for Transformation of Multi Label Classification into Binary Classification

References [1] ,[8] and[11] illustrates that when we write an algorithm for Binary Classification, this algorithm can also be used for Multi Label Classification as follows:-

2.3.1. First Given Condition

It is given that we have a problem of Multi Label Classification under
$\{c_1, c_2, c_3, \dots \dots \dots \dots \dots, c_{|C|}\}$

Where

$C = \{c_1, c_2, c_3, \dots \dots \dots \dots \dots, c_{|C|}\}$: predefined set of categories

2.3.2. Transformation

Transform the problem of Multi Label Classification under $\{c_1, c_2, c_3, \dots \dots \dots \dots \dots, c_{|C|}\}$ into |C| independent problems of Binary Classification under the following set: $\{c_i, \overline{c_i}\}$ for $i = 1\ to\ |C|$

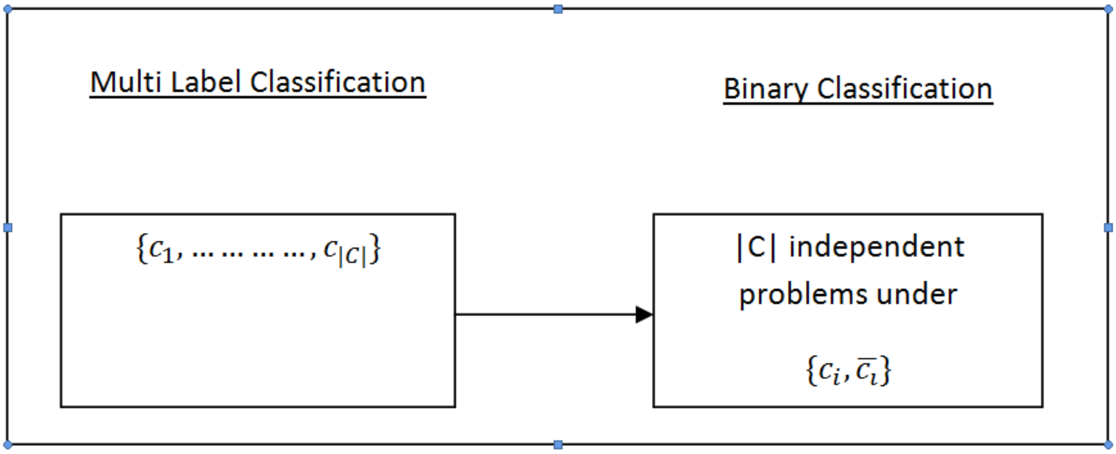

Multi Label Classification
Binary Classification
$\{c_1, \dots\dots\dots\dots, c_{|C|}\}$
|C| independent problems under
$\{c_i, \overline{c_i}\}$

Figure 1. Transformation of Multi Label Classification into Binary Classification

After this transformation we will have a new set of transformed Binary Classification. This new set is following:

$[\{c_1, \overline{c_1}\}, \{c_2, \overline{c_2}\}, \{c_3, \overline{c_3}\}, \dots\dots\dots\dots\dots, \{c_{|C|}, \overline{c_{|C|}}\}]$

2.4. Conditions for Transformation of Multi Label Classification into Binary Classification

For transformation of Multi Label Classification into Binary Classification as discussed in [8] and [11], the following conditions should be met:

-Categories$\{c_1, c_2, c_3, \dots\dots\dots\dots\dots, c_{|C|}\}$ should be stochastically independent of each other.
-For any $\{c', c''\}$ the value of $\Phi(d_j, c')$ does not depend on the value of$\Phi(d_j, c'')$.
-For any $\{c', c''\}$ the value of $\Phi(d_j, c'')$ does not depend on the value of$\Phi(d_j, c')$.

Since we have stated that an algorithm for Binary Classification can also be used for Multi Label Classification but its contrary is not true. An algorithm for Multi Label Classification cannot be used for Binary Classification. Similarly an algorithm for Multi Label Classification cannot be used Single Label Classification.

2.5. Possibilities of Classification of a documentd_j:

The Document Classification is discussed in detail in [12]. It is given that there is a document d_j to classify. There are two possibilities of classification of a documentd_j:

-The classifier might ascribe k > 1 categories to d_j.When k > 1 categories are ascribed tod_j, the question is raised for the selection of “most suitable category” from them. An algorithm for the selection of most appropriate category should be written which should encompass all dependencies of these categories on pertinent factors for document d_j classification.

-The classifier might ascribe no category to d_j.When no category among a set of categories $\{c_1, c_2, c_3, \dots\dots\dots\dots\dots, c_{|C|}\}$ is ascribed to d_j,the question is raised for the selection of "least unsuitable category" from them. When no category is attributed to documentd_j, the least inappropriate category from C should be selected. The question is raised how to select "inappropriate category" from C.

2.6. Significance of Binary Classification

Binary Classification is more useful and significant than Multi Label Classification as illustrated in [9]. The Binary Classification is discussed in more detail in [13].There are various reasons for this significance of Binary Classification.

-There are a number of Text Categorization applications which consists of Binary Classification problems. In fact Binary Classification involves bifurcation of a problem into two sub problems. Each sub problem includes the original sub problem and the complement of original sub problem as constituents of the set. An important Text Categorization application is Filtering. Filtering also comprises of Binary Classification problem which has the fundamental functionality of decision making as illustrated in [9] .For example deciding whether $d_j$is about "Indian political affairs" or not.

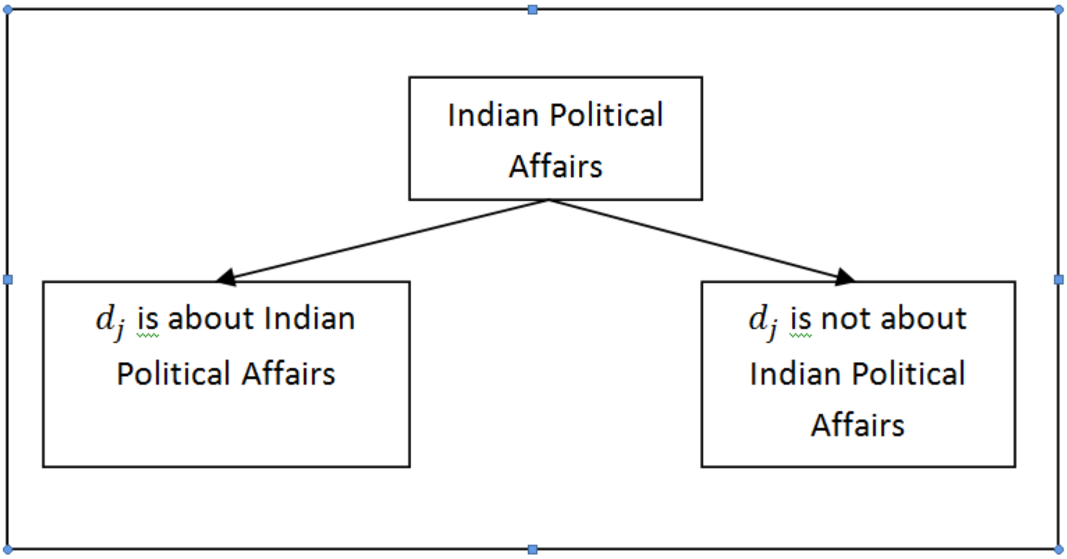

Indian Political Affairs
d_j is about Indian Political Affairs
d_j is not about Indian Political Affairs

Figure 2. Demonstration of filtering which may include decision making

-The most important feature of Text Categorization applications is that most Binary Classification problems include "unevenly populated categories". That is it is always not possible to divide a Binary Classification problem into evenly partitioned categories. Most of the time, Text Categorization problems divide Binary Classification problem into two categories of different population size. For example when we categorize the news about Indian political affairs according to Binary Classification, much fewer documents may be in the class "$d_j$is about Indian political affairs" and more documents may be in the class "$d_j$is not about Indian political affairs".

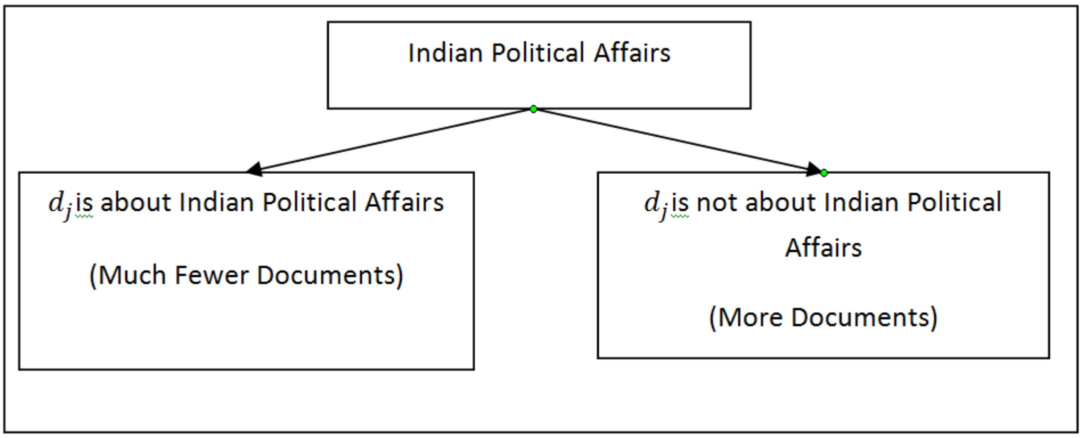

Indian Political Affairs
d_j is about Indian Political Affairs
(Much Fewer Documents)
d_j is not about Indian Political Affairs
(More Documents)

Figure 3. Division of Binary Classification Problem into unevenly populated categories.

-In the same manner, most Binary Classification problem may include "unevenly characterized categories". This means that "what is about Indian political affairs" is characterized better than "what is not about Indian political affairs".

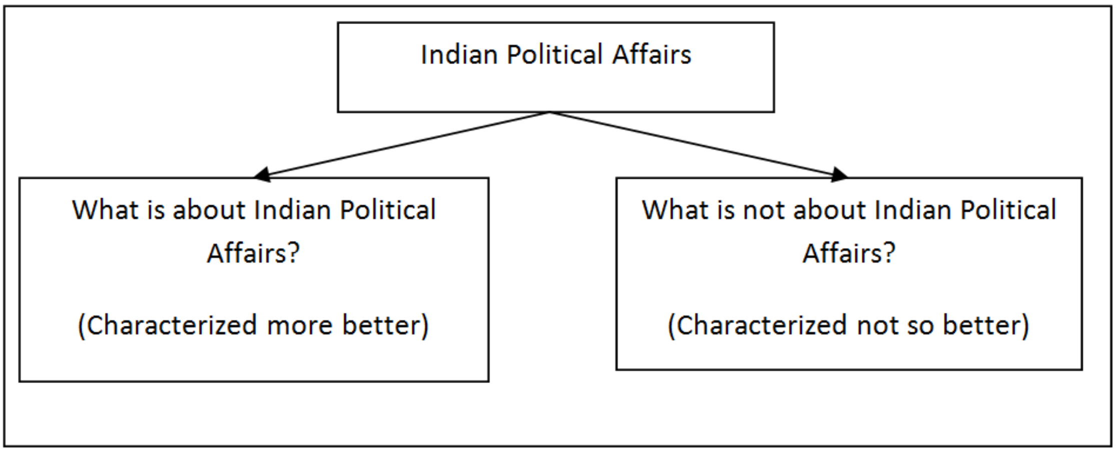

Indian Political Affairs
What is about Indian Political Affairs?
(Characterized more better)
What is not about Indian Political Affairs?
(Characterized not so better)

Figure 4.Division of Binary Classification problem into unevenly characterized categories

-In my opinion, division of Binary Classification problem into unevenly populated categories is according to quantitative evaluation where as division of Binary Classification problem into unevenly characterized categories is according to qualitative evaluation.

-A most common observation in Text Categorization applications is that Binary case is more general than Multi Label case, therefore due to the generalization of Binary case ,when we solve the Binary case this means we solve the Multi Label case. This generalization is implemented in automated indexing for Boolean systems.

-Most of the Text Categorization applications are implemented in binary case.

-It is very easier to understand Binary Classification because most of the techniques for binary classification are just special cases of existing techniques for the single label case.

2.7. Limitations of Single Label and Multi Label Text Categorization

Different scientists had worked on Text Categorization but their work on Text Categorization is limited by certain constraints. For example , consider the case of Single Label and Multi Label Text Categorization .Single Label and Multi Label Text Categorization is implemented for ‘k’ categories that is either less than or equal to ‘k’ or greater than or equal to ‘k’ where ‘k’ is an integer. If the value of integer ‘k’ increases to very larger value, the complexity of the algorithm for implementing Text Categorization will become larger. The interested researchers may work on the complexity of algorithms for implementation of Text Categorization.

3. DOCUMENT PIVOTED VERSUS CATEGORY PIVOTED TEXT CATEGORIZATION

A Text Classifier can be used into two ways. The first one is Document Pivoted Categorization (DPC) and the second one is Category Pivoted Categorization (CPC) as discussed in [14].The basic difference between Document Pivoted Categorization and Category Pivoted Categorization is that document $d_j \in D$is given in Document Pivoted Categorization and category $c_i \in C$ is given in Category Pivoted Categorization. First of all I will define Text Classifier which is also known as Classifier.

3.1. Definition of Text Classifier

The concept of Text Classifier is that there is a document D and we have to classify the document D under a specific given category as discussed in [15]. In totality we have |C| categories and the document D may belong to any one category at a time. If the document D belongs to the specific category then the Text Classifier will be true and if the document D does not belong to the specific category then the Text Classifier will be false. Therefore the Text Classifier is a function which is denoted by a set of two logical values which are true and false.

The mathematical definition of “Text Classifier” is given in the paper referred as [15] below: The classification of text under$C = \{c_1, c_2, c_3, c_4, \dots \dots \dots \dots, c_{|C|}\}$ as consisting of |C| independent problems of classifying the documents in D under a given category c_i for $i = 1\ to\ |C|$.A classifier for c_i is then a functionφ_i:$D \rightarrow \{T, F\}$that approximates an unknown target functionφ_i:$D \rightarrow \{T, F\}$.

3.2. Document Pivoted Categorization (DPC)

It is given that$d_j \in D$, we have to find all $c_i \in C$under which it should be filed. This means that document d_j is searched under all categories and the required corresponding category will be found which contain given documentd_j.

3.3. Category Pivoted Categorization (CPC)

It is given that$c_i \in C$, we have to find all $d_j \in D$under which it should filed. This means that a category $c_i$is searched under all documents and the required corresponding document will be found which contain given categoryc_i.

3.4. Availability of information in sets C and D

When the software or tool incorporating Text Mining principles and techniques is implemented then the theoretical principles cannot be implemented in entirety and certain practical considerations become obvious at the time of implementation which were not considered by Text Mining scientists during hypothesis development. In order to implement Document Pivoted Categorization and Category Pivoted Categorization, two sets are important: set C and set D. It may be possible that these sets C and D may be available at the beginning in entirety or not. It may happen that sets C and D may contain incomplete categories and incomplete documents.

3.5. Effect of Classifier Building Method

The other significant effect is of the Classifier Building Method. We are talking about the Text Classifier Building Method. We have two choices for building Text Classifier. The first choice is Document Pivoted Categorization style and the second choice is Category Pivoted Categorization style. Building text classifiers using either the first style or the second style will have their own effects as illustrated in [16]. From the point of view of ease of use, Document Pivoted Categorization is simpler to implement than Category Pivoted Categorization.

Document Pivoted Categorization is more pertinent when documents are available at different time instants for example filtering an email. The emails related to the social networking are alienated from the official emails. In my opinion web search engines like Yahoo, Google, and Alta vista return the results of their search on the basis of Category Pivoted Categorization. The search engines generate their browsing results by comparing the text typed by the user among a huge collection of categories and finally return results narrowed down by Category Pivoted Categorization.

3.6. When Category Pivoted Categorization should be used?

There are two particular conditions during which Category Pivoted Categorization is used.
-Suppose that we have a set of categories represented as C and a number of documents have been classified under the set of categories C.

$$C = \{c_1, c_2, c_3, \dots \dots \dots \dots \dots, c_{|C|}\}$$

It is required that a new category $c_{|C|+1}$has to be added among the set of categories C then the new set of categories C' will be following:

$$C' = \{c_1, c_2, c_3, \dots \dots \dots \dots \dots, c_{|C|}, c_{|C|+1}\}$$

It is obvious that the addition of new category$c_{|C|+1}$will have an impact on the documents which have already been classified under the set of categories C.

-When the new category$c_{|C|+1}$is added among the set of categories C, the documents should be reconsidered for classification among the new set of categories:

$$C' = \{c_1, c_2, c_3, \dots \dots \dots \dots \dots, c_{|C|}, c_{|C|+1}\}$$

Therefore we can say that Category Pivoted Categorization is more typical and complex than Document Pivoted Categorization. Category Pivoted Categorization has more burden of work for recurrent classification of documents when a new category is added in the given set of categories. Document Pivoted Categorization is more commonly used than Category Pivoted Categorization because of the simplicity and applicability of the approach.

4. Hard Categorization Versus Ranking Categorization

When we talk about Text Categorization, it necessarily involves the automation of Text Categorization. The automation of Text Categorization is either complete or partial. When Text Categorization is fully automated then a decision of either true or false is taken for each pair of the following tuple:

$$\langle d_j , c_i \rangle$$

The term automation means involvement of human expert is minimized and the software which performs Text Categorization operation is intelligent enough to take decision about each pair $\langle d_j , c_i \rangle$.The Full Automation of Text Categorization requires True, False decision on each pair of document and category as illustrated in [18]. Let us consider an example. Suppose that there is a document $d_1$belongs to the category c_2 then for the pair of $< d_1, c_2 >$.Full Automation will yield True decision. Consider another example, if a document $d_4$does not belong to the categoryc_8, then for the pair of$< d_4, c_8 >$, Full Automation will yield false decision. When we talk about the actual implementation of Full Automation of Text Categorization, we need specific tables on Data Servers containing three fields: Documents, Category and Belonging. Consider the table below:

Table 1. Full Automation of Text Categorization.

Document	Category	Belonging
d_1	c_2	True
d_4	c_8	False
d_2	c_1	True
d_2	c_3	False
d_5	c_8	True
d_6	c_2	True
d_7	c_2	True

If partial automation of Text Categorization is performed then the requirements and needs will be very different. For example if it is given that$d_j \in D$, we have a set of given categories

$C = \{c_1, c_2, c_3, \dots\dots\dots\dots\dots, c_{|C|}\}$.In the given set of categories C, the system can do the ranking among categories as illustrated in [19]. The method to perform ranking among categories is that the system calculates the “estimated appropriateness” of d_j.No hard decision is taken on any category. It is obvious that after ranking is performed the final list will be very changed. The final list obtained from ranking will be in a specific order. This final list will be called Ranked List. Therefore, we can define Ranked List as a list of categories which is sorted according to estimated appropriateness of d_j.The finally obtained Ranked List is very important and helpful from the point of view of human expert. The reason is that the human expert is the key person responsible for taking the final categorization decision. The human expert will sort the categories from the top of the list only because he will restrict the choices at the top of the list only to the categories. The advantage obtained from the automated ranked list is that there is no need to examine the entire list; only top of the list will be used by the human expert to find categories with estimated appropriateness to d_j.

For example suppose that the set of categories initially available for ranking is:

$$C = \{c_1, c_2, c_3, c_4, \dots\dots\dots\dots\dots, c_{|C|}\}$$

Suppose that c_4 has the higher rank in terms of estimated appropriateness to d_j as compared to c_1, c_2, c_3.Then the final ranked list C_R will be following:

$$C_R = \{c_4, c_1, c_2, c_3, \dots\dots\dots\dots\dots, c_{|C|}\}$$

Further investigation into the Partial Automation of Text Categorization tasks introduces the concept of Fuzzy Logic along with estimated appropriateness to d_j. We can use Fuzzy Logic and Fuzzy sub set values that is semi true, semi false decisions on each pair of $\langle d_j, c_i \rangle$ along with estimated appropriateness to d_j which may produce better belonging of a document $d_j$to category c_i and it will give very good results as compared to the techniques discussed and illustrated in [7].The technical justification of using Fuzzy Sub Set Values is that during the Text Categorization Task we may deal with a document d_j containing text which partially belongs to a category c_i.Again there is a role of human expert who will take the decision about the partial truth or partial falseness of the belonging of document $d_j$to a category c_i.The future research direction can be the development of an algorithm for finding Fuzzy Sub Set Values for representing the belonging of document $d_j$to a category c_i in Partial Automation of Text Categorization.

4.1. Some Examples of Partial Automation of Text Categorization

Let us consider the following text from the Daily Dawn News Paper of Pakistan.
“With a gap of over 17 years, the government on Saturday gave final touches to arrangements for carrying out population and housing census in March this year. Preliminary results will be compiled within three months. The governing council of Pakistan Bureau of Statistics (PBS) Headed by Finance Minister Ishaq Dar approved a revised time line for the census. The preliminary results will have to be completed by June this year, while other aspects including district wise data reports will have to be completed by December 2017.An official source privy to the meeting told Dawn that although compilation of population reports after census normally took around three years, the government wanted to be completed it before the next general elections due in 2018.The compilation of all reports by December 2017 will give enough time to political parties and the Election Commission to use data for electioneering purposes. At the moment there are a number of bills pending before both houses of parliament seeking increase in the number of seats for minorities as well as representation of provinces, especially Baluchistan in the National Assembly. While taking up those bills, the committees of the two houses had already decided to defer consideration of the bill until the finalisation of the census data. The population census will yield statistics about internal migration, urbanisation as well as urban and rural population across the country. The population data will be used for delimitation of the constituencies of the national and provincial assemblies, a requirement under the constitution. According to the source, the exact dates have not yet been finalised .However; it has been decided to hold census at the end of

March .The budget approved for the census is Rupees 4.5 billion, which will be shared by the provinces. The population census will be conducted; the first three days for a house listing operation and the following 15 days for the main count which include filing of census forms on house to house level and a day for homeless people. The PBS says that the census will be held with the full support of armed forces at a man to man level as was done in 1998.An official statement issued after the meeting by the finance minister said that Finance Minister Ishaq Dar at the outset welcomed the newly appointed PBS members and mentioned that they had joined the organisation at the crucial time, just a couple of months before the sixth population and housing census. They would have to shoulder the responsibility of making the census successful, he said. Chief Statistician Asif Bajwa informed the meeting that as per the decision of the governing council, PBS has adjusted the target dates for preparation of census reports by December 2017.He said that the PBS had initiated all preparations and held meetings with provincial authorities to brief them about their responsibility in making the census successful and transparent. Mr. Dar said it is the prime responsibility of PBS to conduct credible census. He said that utmost care should be taken to complete the task in a transparent manner for a credible data. Such data he said could then form the basis for the future planning. During the meeting it was agreed that qualified and highly reputed statisticians would be co-opted as members of the technical committee to benefit from their expertise. Mr. Dar directed that frequent meetings of the technical committee be held until the census was held and suggestions of experts be brought to the governing council for consideration."

This article was published on 3rd of January 2016 under the heading "Population census to be held in March" by Mubarak Zeb Khan. We will perform Partial Automation of Text Categorization for this article published in the Daily Dawn.

The idea was discussed in [7].Let us suppose that

$c_1 \in{}' \mathit{PopulationCensus\ to\ be\ held\ in\ March}'$

$c_2 \in{}' \mathit{Population\ Data\ for\ the\ use\ of\ Electioneering\ Purpose}'$

$c_3 \in{}' \mathit{Effortsof\ PBS\ for\ successful\ organization\ of\ census\ in\ Pakistan}'$

$c_4 \in{}' \mathit{StatisticalData\ of\ Census}\ '$

$c_5 \in{}' \mathit{Roleof\ reputed\ statisticians\ in\ holding\ successful\ census}\ '$

Table 2. Partial Automation of Text Categorization.

Document	**Category**	**Estimated Appropriateness to Document**
d_1	c_2	0.6
d_1	c_4	0.4
d_1	c_1	1.0
d_1	c_3	0.5
d_1	c_5	0.2

The final Ranked List for the above example according to estimated appropriateness tod_j will be following:

$C_R = \{c_1, c_2, c_3, c_4, c_5\}$

The final Ranked List will be used by the human expert for Text Categorization using Partial Automation of Text Categorization method.

4.2. Semi Automated Classification Systems

Semi Automated Classification System is also known as “Interactive” Classification System. Semi Automated Classification Systems are discussed in detail in [20]. When we compare human expert with fully automated system then the effectiveness of human expert is very larger than the fully automated system, therefore semi automated systems are very useful and effective. The reason for effectiveness of semi automated systems is that they incorporate advice of human expert in addition with the results generated by automated systems. When we design Automated Systems or Semi Automated Systems, we cannot deny the significance of training data and training documents. Training documents are used when actual documents needed for classification are not available and Training data is used when actual data needed for classification is not available. Semi automated systems are useful when the following conditions are fulfilled:

-When quality of training data is less.

-When the documents that have to come for automated classification systems are not known and not seen then the training documents are used for semi automated classification system but the training documents cannot be trusted to be representative sample document of the unseen documents that have to be classified using automated classification system.

Therefore we can conclude that semi automated classifiers are more effective and useful than fully automated classifier and the results of a fully automated classifier cannot be trusted completely.

4.3. Future Research Directions

In Text Categorization it is quite obvious from the mathematical definitions of Text Categorization that categories are assigned to the documents. For example we have a paragraph containing the text from the political affairs of Pakistan, and then this paragraph of text will be covered under the category of ‘Politics’. If we have a paragraph containing text from the sports news of Pakistan, then this paragraph of text will be covered under the category of ‘Sports’. Suppose that we have three paragraphs of texts, the first paragraph of text is covered under the category ‘Pakistani Politics’, the second paragraph of text is covered under the category ‘Indian Politics’ and the third paragraph of text is covered under the category ‘Bangladeshi Politics’, then the whole three paragraphs will be categorized under single category ‘Politics’. In this example we have three Multi Label Text Categorizations and one Single Label Text Categorization. In other words we have three Multi Label Text Categorizations under one Single Label Text Categorization. When we move from one Single Label Text Categorization to three or more Multi Label Text Categorizations, we need proper mathematical models and algorithms for grouping many Multi Label Text Categorizations under one Single Label Text Categorization. Also we can calculate complexity of algorithm for grouping many Multi Label Text Categorizations under one Single Label Text Categorization.

The next future research direction is obtained from the above discussion of possibilities of classification of documentd_j. The research direction is that we have found that the set of categories$\{c_1, c_2, c_3, \dots\dots\dots\dots\dots, c_{|C|}\}$ should be partitioned into subset of categories according to their appropriateness. An interested researcher on Text Mining may work on this problem for finding an algorithm which will partition predefined set of categories $\{c_1, c_2, c_3, \dots\dots\dots\dots\dots, c_{|C|}\}$ into subsets according to appropriateness.

5. CONCLUSIONS

In this paper our main focus is to classify a document into a range of categories. We have given a set of categories and the document may belong to any one category at a time. The relation and association of the document to a specific category is defined by a logical function which is a Boolean function here .This Boolean function is called Text Classifier and the purpose of Text Classifier is to define the logical state of the belonging or relation of the document to a specific category. In this paper we have compared the Semi-Automated Classifiers with Fully-Automated Classifiers. After comparison of Semi-Automated Classifiers with Fully-Automated Classifiers we found that Semi-Automated Classifiers are quite better than Fully-Automated Classifiers. Fully-Automated Classifiers should not be used and are not trustworthy.

ACKNOWLEDGEMENTS

I would like to thank the ex- 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. Also I would like to thank Dr. Farah Naz Lakhani for motivating me to contribute towards research in the field of Computer Systems Engineering and Computing.

REFERENCES

[1] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," Knowledge-Based Systems, vol. 24, pp. 1024-1032, 2011.

[2] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TF* IDF, LSI and multi-words for text classification," Expert Systems with Applications, vol. 38, pp. 2758-2765, 2011.

[3] S. Jiang, G. Pang, M. Wu, and L. Kuang, "An improved K-nearest-neighbor algorithm for text categorization," Expert Systems with Applications, vol. 39, pp. 1503-1509, 2012.

[4] S. Wang, D. Li, X. Song, Y. Wei, and H. Li, "A feature selection method based on improved fisher's discriminant ratio for text sentiment classification," Expert Systems with Applications, vol. 38, pp. 8696-8702, 2011.

[5] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," Knowledge-Based Systems, vol. 36, pp. 226-235, 2012.

[6] M. L. Jockers and D. M. Witten, "A comparative study of machine learning methods for authorship attribution," Literary and Linguistic Computing, p. fqq001, 2010.

[7] J. Yang, Y. Liu, X. Zhu, Z. Liu, and X. Zhang, "A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization," Information Processing & Management, vol. 48, pp. 741-754, 2012.

[8] J.-Y. Jiang, S.-C. Tsai, and S.-J. Lee, "FSKNN: multi-label text categorization based on fuzzy similarity and k nearest neighbors," Expert Systems with Applications, vol. 39, pp. 2813-2821, 2012.

[9] C. Jiang, F. Coenen, R. Sanderson, and M. Zito, "Text classification using graph mining-based feature extraction," Knowledge-Based Systems, vol. 23, pp. 302-308, 2010.

[10] Z. Li, Z. Xiong, Y. Zhang, C. Liu, and K. Li, "Fast text categorization using concise semantic analysis," Pattern Recognition Letters, vol. 32, pp. 441-448, 2011.

[11] Z. Kalal, J. Matas, and K. Mikolajczyk, "Pn learning: Bootstrapping binary classifiers by structural constraints," in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, 2010, pp. 49-56.

[12] V. Srikumar and C. D. Manning, "Learning distributed representations for structured output prediction," in Advances in Neural Information Processing Systems, 2014, pp. 3266-3274.

[13] A. Sarker and G. Gonzalez, "Portable automatic text classification for adverse drug reaction detection via multi-corpus training," Journal of biomedical informatics, vol. 53, pp. 196-207, 2015.

[14] B. Mitra, E. Nalisnick, N. Craswell, and R. Caruana, "A Dual Embedding Space Model for Document Ranking," arXiv preprint arXiv:1602.01137, 2016.

[15] S. T. Dumais, D. Heckerman, E. Horvitz, J. C. Platt, and M. Sahami, "Methods and apparatus for classifying text and for building a text classifier," ed: Google Patents, 2001.

[16] E. Cambria and B. White, "Jumping NLP curves: a review of natural language processing research [review article]," Computational Intelligence Magazine, IEEE, vol. 9, pp. 48-57, 2014.

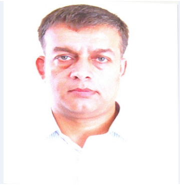
[17] Y. Aphinyanaphongs, L. D. Fu, Z. Li, E. R. Peskin, E. Efstathiadis, C. F. Aliferis, et al., "A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization," Journal of the Association for Information Science and Technology, vol. 65, pp. 1964-1987, 2014.

[18] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic text categorization in terms of genre and author," Computational linguistics, vol. 26, pp. 471-495, 2000.

[19] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Computers & Electrical Engineering, vol. 40, pp. 16-28, 2014.

[20] G. Berardi, "Semi-automated text classification," in ACM SIGIR Forum, 2014, pp. 42-42.

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, Text Mining, Parallel Processing, CAO, and Statistical Learning.