RESEARCH ARTICLE OPEN ACCESS

Enhancing Dependency using Power Set Theory for Text Classification

DoaaMabrouk^a, SherineRady^b,NagwaBadr^c and 〖M.E.Khalifa〗^d
(_^(a,d))Cairo,Egypt,Faculty of Engineering,Egyptian Chinese University.
(_^(b,c))Cairo,Egypt,Faculty of Computer and information Sciences,Ain Shams University.
Corresponding Author: Doaamabrouk

ABSTRACT
There are many models to solve Information Retrieval problems. Some of them assumeterms are independent while othersassume terms are dependent.Models that assume independence, may lead to synonymy and polysemy problems. Synonymy occurs when using different terms for the same concept, while polysemy is concerned with ambiguity of the term.Text classification is an important aspect in different areas. Text Classification is based on matching the words in different documents and retrieving class labels. It may be applied either manually or automatically. In this paperweuse the power set theory to enhance the dependency in Text Classification using discovering unique words and "Term Dependency Identification" that distinguishes different documents from one another. The classical dataset was used to find Term Dependency Identification and measure accuracy through Subsumption Rule Based Classifiers into two different ways: Maximum-Number-Term Dependency Identification and Maximum-Feature Count. The 5-fold Cross Validation Experiments give results with average accuracy 96%.
Keywords: Text Classification, term dependency, neural network, deep learning and Power set

Date Of Submission: 09-05-2019 Date Of Acceptance: 24-05-2019

I. INTRODUCTION

Dependency is one of the problems in IR. It means finding words that are related to each other. Dependency relation may be linguistic or grammatical. In grammatical dependency, the relation is between pairs of terms while in linguistic dependency the relationis considered a binary relationship between words. The best way for representing dependency (is when) "X->Y"this means that "Y" depends on "X". When the linguists talked about dependency, they probably often talked about "syntactic dependency". This leads to logical mistakes because there are natural language features that have at least three types of dependency. these types are semantic, syntactic and morphological.Some models which assume terms are independent,made retrieval easier to be implemented.

There are many researches that have been carried out in developing the term dependency in Text Classification (TC). TC helps inclassifying documents into sets of predefined categories. There are many algorithms used in TC such as K-nearest neighbor (KNN), naïve Bayes Classifiers (NBC), Decision Tree (DT), Support Vector Machine (SVM) and Artificial Neural Network (ANN).These algorithms, that have been used, offer acceptance performance, but some of them are very costly such as naïve Bayes Classifiers (NBC), Vector Space Machine (VSM), Copula language model in [12]. SVM produce betterclassification than NBC. However, NBC is still widely used because it is simple and efficient. Designing algorithms that model dependency between terms in each document for each category can improve retrieval and effectiveness.Text classification (TC) [9,7], is one of the most important tasks in natural language processing (NLP). TC is an example of a supervised machine learning taskso thelabelled dataset contains text documents was used. When using TC, remove non-informative terms to improve effectiveness and reduce computation time.The goal of using text classification in dependence problem is to make retrieval easier and give the exact solution through retrieving class label.

In this paper, wepropose the power set theory to enhance dependency in text classification. The possible advantage is to find all possible combinations between terms and unique terms in the documents. These terms are distinguished documents from each other. On the other hand, it is the first time to apply such algorithm in the retrieval field. One of the difficulties of using power set is time consumptionand platform. A series of experiments on standard classical dataset

have been conducted to find term dependency identification for each document in each category. Moreover, Subsumption Rule Based Classifiers were used to evaluate the results accuracy.

II. RELATED WORK

In [3], a novel dependency model to combine word net and co-occurrence relationship in language model for information retrieval was used. This model was applied on TREC collections. In case of word net, it was used to cover terms related to each other that cannot beidentified automatically.The shortcoming isthere is no associated weight, and also the relation between terms isbinary. This model is concerned with the second type of dependency "Syntactic",whilethe relation is classified as synonymy, hypernymy and hyponym.Link model (LM), non-separated link model (NSLM) and unigram were compared. LM, NSLM are outperformed on three datasets. In retrieval, LM is better than unigram because it depends on direct match between the document and query. If LM is used, the result is not good and it is very low in terms of performance as it works only on 10 Mega.In case of a combination between unigram and co-occurrence (CM), the result is good and CM works on 40 times over LM. Term Relevance dependency model (TRDM) for text classification (TC)was introduced in [2]. Integrating relationship between words is very important and has a great interest. The unigram LM for TC is based on matchingthe literal words in the documents and classesto capture the semantic relationshipsof words in IR.This model was applied in 20newsgroups and reuters-21578. The novel model outperforms the standard NBC and several LM-NBC based TC. Syntactically related word pairs using dependency parser werepresented in [15]. This model is applied in Reuters dataset. The result of the proposed model was that the new model is better in terms of accuracy and precisionwhetherit was used single or combined with unigram model rather than using unigram solely.A Concept-Based language model was presented in [6]. Many models of IR assume terms are independent. These models demonstrateda good performance. LM is concerned with dependency such as bigram, concept-phrase or word relationship. LM performs better compared to traditional IR models such as VSM and probabilistic. Concept based LM assume concept might be a single word or multiple words and might be an ontology or a frequent collection in documents. This model was applied on TREC collection. The new model achieved improvement over MRF and unigram model in terms of mean average precision "MAP".Novel hybrid dependency structure for describing dependencies between terms was presented in [1]. It allows integration of various forms of dependency with single frameworks. Dependency can be understoodfrom two point of views. Firstly, dependency between terms within query or within document. Secondly, dependency between query terms and document terms. This model focusses on the first definition of dependency. There are two problemsthat should be taken into consideration when concernedwith term dependency; how to define dependency and how to apply dependency between terms in retrieval. Earlier term dependency in LM used to capture dependency by bigram or trigram. These models assume dependency between adjacent terms. Dependency within sentences differ and can be described from different perspectives: direct or indirect syntactic relation and proximity relation. In this paper,the researchersuse an intuitive way to define dependency such as syntax based, proximity and co-occurrence-basedways. intuitive way was applied in TRECdisk 4 and 5 for experiment. intuitive wayevaluates methods on ad hoc for TREC6 with topics 301-350, TREC 7 with 351-400 and TREC8 with 401 -450. The proposed model outperformedthe traditional models and improved effectiveness. New fuzzy logic based ranking function (fuzzy inference system "FIS") was proposed in [17] to enhance retrieval system. The ranking function is based on computation of different terms of term weighting such as (TF, IDF, normalization). The computation methods were used to retrieve relationsbetween query and documents.Fuzzy logic wasused at two levels to compute relevance score of documents. The first level consisted of two fuzzy logic controllers (FLC). It contained two parts, one for structuring the feature of document and the other for structuring of queries. The second level consisted of one level (FLC). This method was applied on a classical dataset (CACM, CISI). The result of the proposed work (new ranking function "FIS") wasbetter than that of the fuzzy logic based ranking function. FIS improved the performance of IR. The ranking function increased the value of precision, recall and F-measure.In [11], concept coupling relationship analysis model was proposed to learn and aggregate the intra, inter concept coupling relationship. The classical IR relied on keyword-matching to index document, where queries and documents wererepresented by Boolean, VSM and probabilistic models. The existing retrieval system often returned inaccurate and incomplete results because of the semantic challenges such as polysemy and synonymy. There are various efforts that have been made to address the concept-lattice

based retrieval methods for query transformation. The concept-lattice based retrieval methods expanded, refined query and explored navigation search strategies using specificity or generality relation.Query expansion generated a novel query by augmenting original query with new features with similar measuring. These wereapplied on classical datasets (MED, CACM, CISI, CRAN). The result of the proposed method was achieving improvement in the mean average precision (MAP) with 9%, interpolated average precision (IAP) with 8% and precision with 15%.

Deep learning and neural networks are the most recent research area. These techniques are applied in modern IR systems. These areas can be available for graduates and post graduates. Neural networks are applied in all key points of modern IR such as (Ranking Algorithm), click models knowledge graphs, text similarity language modeling and question answering as in [13]. When training data is available as raw input date, neural network can be applied.Application of recurrent neural network (RNN) such as production, prediction or recognition requires system that will store and update information. The Cycle of RNN graph keeps information about past input for an amount of time. The time is not fixed as it depends on the weight of input data. To store relevant information based on dynamic systems, there are three requirementssuch as store information for despotic duration, resistant to noise and parameters be trainable.Gradient descent technique was presented in [18]. It became increasingly inefficient when temporal span of dependencies increased. Short dependencies weresufficient rather than long term dependencies.If we start training with a short sequence, the system rapidly settles in the correct region of parameter space. RNN is very powerful to represent context and it outperformed static networks.Gradient descent with nonlinear autoregressive models in [14] with exogenous (NARX) wasmore effectivethan RNN. The result of NARX with RNN improved performance on thelong-term dependency and Retain information for as two to three times long as the conventional RNN.A special kind of RNN is concerned with learning long term dependencies in [4]. It wasdesigned to avoid long term dependency problems. The structure of long short-term memory (LSTM) and RNN as the same but RNN differed in the middle layer. Instead of having a single neural layer in LSTM, there werefour interacting with each other in a special way.Biologically inspired deep network (shuttle net) was presented in [16]. Shuttle net consisted of several processors such asGated Recurrent Unit (GRU). GRU is associated with multiple groups of hidden states unlike RNN. It wasapplied in two benchmark datasetsHMDB 51 large collection of realistic videos from different sources including movies and web videos, and UCF 101 most popular action recognition benchmarks. The new technique outperformed LSTM and GUR,despitehaving the same number of parameters.In [10], a novel hybrid text classification model based on deep belief network and softmax regression was introduced. This model is presented to solve high dimensional computation problem. It is applied on REUTERS 21578 and 20 newsgroup datasets. This model is compared with classical models such as Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) and the result is the newest model is outperformed rather than classical model. At the beginning of 1960s, the specialized researchers in neural networks were interested in shallow structure such as a single nonlinear layer that were suitable for small dataset size. But over time and increased size, this model was no-longer optimal and the most appropriate model was deep neural network. The comparative has been done between KNN, SVM, DBN and softmax through accuracy. These comparatives have been made using small and large data size. The result is DBN is the highest accuracy with 82.5% for small data size and 86.66% for large data size.

III. POWER SET THEORY

It is a set of all possible subsets of the set. It is a machine learning algorithm, which can be embedded in the process of data preprocessing, learning and reasoning. It is used to express uncertainty by means of boundary region of a set. Power set theory operates on $\mathbf{2^n}$ features. Where "n" means the number of words in a document, but in this dataset, power set works on total size of documents multiplied bythe maximum size of files. Nonetheless,this is completelyimpossible because of the large document size in the total number of words. In this case,the level set of power set was used. Thus, the best one is power set. The advantages of the power set are giving the exact solution and being concerned with the meaning, while the disadvantage istaking much time. Power set is the best choice rather than heuristic algorithms. Heuristic is used for finding solutions among all the possible ones, but with noguarantee that the best will be found. Accordingly, it may be considered as an approximate way and not accurate. It also fails to find the exact solution. However, it takes less time in solving problems, and gives the optimal solution but not the exact solution. For example,from applied dataset:

Let the set S = {Number, System, Binary, Tree}.The power set of the set S is$\mathbf{2^4 = 16}$. Where

4 is the number of elements in the set (Number, System, Binary, Tree).

3.1 The proposed algorithm

3.1.1 Power set algorithm

OpenKnowledgeBase (Original Knowledgebase)
ClearSet (SetOfEvaluatedFindingsSubSets) /*emptySetOfEvaluatedFindingsSubSets */
ForeachTDI**in** OriginalKnowledgeBase **do**
ClearSet (FindingsSet) /* empty FindingsSet */
FindingsSet = **GetFindings** (Rule) /* get all findings of the specific
TDI in the original knowledge base and put them in FindingsSet */
ClearSet (PowerSetOfFindingsSet)/* empty PowerSetOfFindingsSet */
PowerSetOfFindingsSet = **PowerSet** (FindingsSet) /* returns all sets of power
set of FindingsSetfor specific TDI and put them in PowerSetOfFindingsSet */
DeleteEmptySetFromPowerSet (PowerSetOfFindingsSet)
For each FindingsSubSet **in** PowerSetOfFindingsSet **do**
ClearSet (TDIsIdSet) /* empty TDIsIdSet */
/* to ensure that there is no any subset in SetOfEvaluatedFindingsSubSets
may causesubsumption with FindingsSubSet */
If not (**SubSumptionExists** (FindingsSubSet, SetOfEvaluatedFindingsSubSets))
then
ClearSet (DisordersList)
TDIsIdsSet = **GetRulesIds** (FindingsSubSet) /*Find all TDIs' ids
which have all elements of that set as findings */
For eachTDIId **in**TDIsIdsSet **do**
DisorderName = **GetDisorder** (TDIId) /*return all disorders
name of the specific TDI which has given RuleId*/
AddToSet (DisorderName, DisordersList)
NextTDIId
/*If all elements of the DisordersList are identical */
If (**IsIdenticalList** (DisordersList)) **then**
/* Construct a new TDI and add it to the knowledge base */
NewTDI = **ConstructNewRule** (DisordersName, FindingsSubSet)
AddTDI (NewTDI, RefinedKnowledgeBase)
End if
AddToSet (FindingsSubSet, SetOfEvaluatedFindingsSubSets)
End if
Next FindingsSubSet
Next**TDI**

3.1.2 Subsumption Rule-Based Classifiers (SRBC) algorithm

SubSumptionExists (Findings Subset**,** SetOfEvaluatedFindingsSubSets**)**
PowerSetOfFindingsSubSet = **PowerSet** (FindingsSubSet)/* returns all sets of power set of FindingsSubSet for specific Rule and put them in PowerSetOfFindingsSubSet */
DeleteEmptySetFromPowerSet (PowerSetOfFindingsSubSet)
For each PowerSetElement **in**PowerSetOfFindingsSubSet **do**
For each EvaluatedElement**in**SetOfEvaluatedFindingsSubSets **do**
If subset (EvaluatedElement, PowerSetElement) **then**
SubSumsionExists = True
Exit
End If
NextEvaluatedElement
NextPowerSetElement
IsSubSumsionExist = **False**

IV. EXPERIMENT

4.1 Dataset Description

One well benchmark dataset used in IR is a classic collection dataset. This dataset consists of four different class labels (categories). It is about different fields such as Medical (Med), Character Institute for Securities and Investments (CISI), California Association of Community Managers (CACM) and CRANFIELD. The content of each categoryismedical articles, articles about information sciences, articles from communications of the ACM journal and abstracts from aeronautics articles respectively. The size of it is nearly 7019 documents in all categories. Each category has a different number of documents and each document is different in size (Length) also which makes it the best choice.Table 1. Shows the total number of documents in training and others for testing with different percentage. Table 2. Shows the total number of the words in each category and the average words in training and testing. This means that the p-value and t-value are equal.This shows a relationship between the sample and population.

Table 1. Training and testing of all datasets for experimentation.

Dataset	20%	80% training
	Number of testing	Number of training
CACM	617	2568
CRAN	253	1128
MED	171	843
CISI	266	1173
Total	1307	5712

Table 2. Average words in each category.

Dataset	20%	80% training
	Total words in testing	Total words in training
CACM	20575	86138
CRAN	68312	295324
MED	85231	377245
CISI	41640	181599
Total	215758	940306
Average total	165	165

V. EVALUATION CRITERIA

The evaluation of power set algorithm was carried out by applying it to classical datasetswhich consisted of four labels. This has already been developed using visual studio (c#) version 2017 and the hardware wasa server. The server specification was about 128 GIGA RAM, windows server 2017 and 1 TERA Solid State Drive (SSD). Table 3. Shows information about 5-fold cross validation experiments results.Conversion of the dataset into XML format before power set algorithm was applied on them. The results of this algorithm contain uniqueness of each training and size on disk in Kilobytes (KB). Table 4. displays the distribution of documents in each category in minimum and maximum TDI and the range between them.The subsumptionrule-based classifiers "SRBC" was used to evaluate the accuracy of the output result coming from power set theory. This technique was presented in [8].It contained two ways maximum- number- term dependency identification "Max-No-TDI" and maximum feature count "Max-FC". The equation that calculated accuracy was as follows:

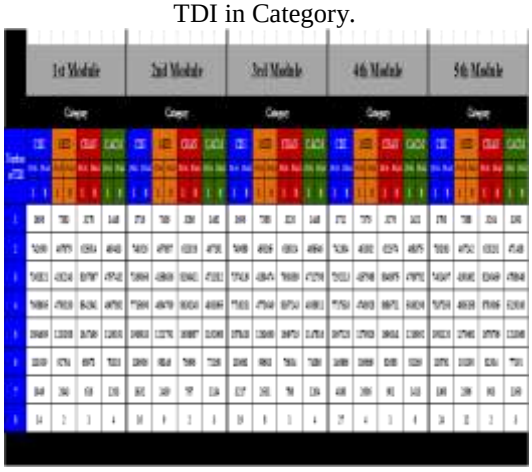
$$\text{Acc} = \frac{N_c}{N_t} \times 100 \qquad (1)$$

WhereN_c: Number of corrected instances that included a large number ofTDIs for max-no rule.N_t: Total number of the all TDIs.

Table 3. Information about 5-fold cross validation experiments

No.of Experiments	Searched objects	Uniqueness	Size on disk (KB)
EXP.1	5712	58085056	4860
EXP.2	5712	58031881	4829
EXP.3	5712	58017328	4811
EXP.4	5712	58437742	4831
EXP.5	5712	59413234	4870

Table 4. Distribution of Minimum and Maximum TDI in Category.



VI. RESULTS AND DISCUSSION

The result is XML format called "Module". The maximum and maximum numbers of TDI were calculated for each module. Table 4shows the analysis of the minimum and maximum TDI values. The minimum number of TDI was"1" while the maximum was"8" for 5-fold cross validation experiments. The total minimum number of TDI in the minimum TDI was"3220", while the maximum number in the minimum was"7405". The total minimum number of TDI in the maximum was"1", while the total maximum in the maximum was"27". Figure 1 shows the minimum and maximum TDI values for each module. This module wasan input and testing file wasused for evaluating this module through SRBC including two criteria Max-No-TDI and Max-FC. This step wasrepeated for all modules that came out from the power set theory. Table 5shows the result of the accuracy and Figure 2 shows the percentage. From this analysis, Max-No-TDI gavehigh results rather than Max-Fc.

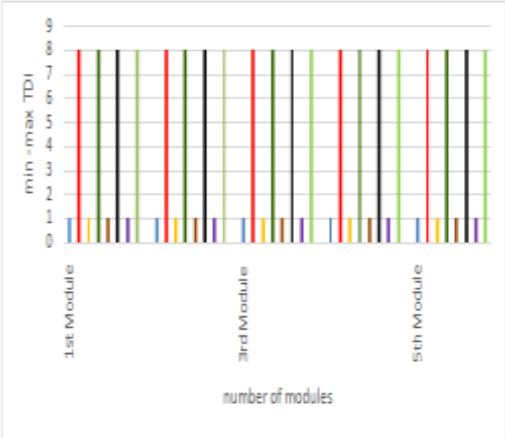


Figure 1. Minimum and Maximum TDI.

Table 5. The accuracy of dependence results from the power set.

	EXP.1		EXP.2		EXP.3		EXP.4		EXP.5	
Experiments	Max no rule	Max-fc	Max no rule	Max -fc	Max no rule	Max-fc	Max no rule	Max-fc	Max no rule	Max-fc
Accuracy	96%	84 %	83.15 %	24.44%	84.02 %	23.45 %	83%	26.1 %	83.51 %	24.44 %

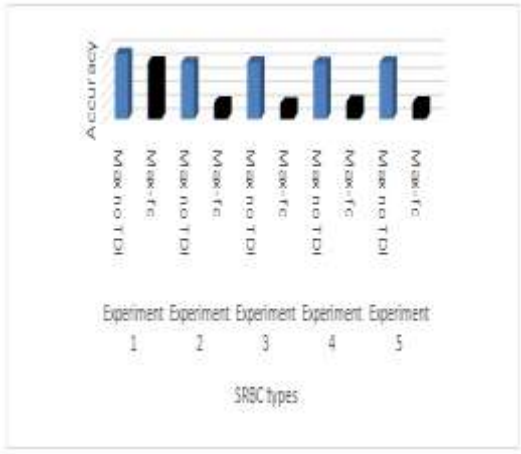


Figure 2. The percentage of the accuracy.

VII. CONCLUSION

In thispaperwepresented the power set theory to find dependency between terms in each document for each category.At the beginning, weperformedpreprocessing for the dataset to extract all features of the documents, separatedthe features with commas and ended them with the class label. These documents wereunstructured data, thus, we developeda tool to convert these documents into XML. At the end, the algorithm wasapplied and its accuracy was measured as well.

List of Figures

No	Figure Name
1	Minimum and Maximum TDI
2	The Percentage of the accuracy

List of Tables

No	Table Name
1	Training and Testing of all datasets for experimentation
2	Average Words in each Category
3	Information about 5-fold Cross Validation Experiments
4	Distribution of Minimum and Maximum TDI in Category
5	The Accuracy of Dependence Results from the Power Set

REFERENCES

[1]. CAI Ke-ke, BU Jia-jun, CHEN Chun, QIU Guang; "A novel dependency language model for information retrieval"; Journal of Zhejiang University SCIENCE A ISSN 1673-565X (Print); ISSN 1862-1775 (Online),2007.

[2]. Donald H. Kraft, Erin Colvin; "Fuzzy Information Retrieval"; 2015.

[3]. Guihong Cao, Jian-Yun Nie and Jing Bai; "Integrating Word Relationships into Language Models"; SIGIR'05, August 15–19, 2005, Salvador, Brazil, Copyright 2005 ACM.

[4]. Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, Rabab Ward; "Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval"; Transactions on Audio, Speech, and Language Processing; IEEE; 2015.

[5]. Hang Li, Zhengdong Lu; "Deep Learning for Information Retrieval"; SIGIR; 2016.

[6]. Lynda Said Lhadj, Mohand Boughanem, Karima Amrouche; "Enhancing Information Retrieval Through Concept-Based Language Modeling and Semantic Smoothing"; JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY, 2016.

[7]. Maher Abdullah and Mohammed GH. I. AL ZAMIL; "The Effectiveness of Classification on Information Retrieval System (Case Study)"; 2018.

[8]. Mahmoud Nasr, Khaled El-Bahnasy, M. Hamdy, and Sanaa M. Kamal; "A Novel Model based on Non Invasive Methods for Prediction of Liver Fibrosis"; IEEE; 2017.

[9]. Meng-Sung Wu, and Hsin-Min Wang; "Term Relevance Dependency Model for Text Classification"; 21st International Conference on Pattern Recognition (ICPR 2012)November 11-15, 2012. Tsukuba, Japan

[10]. MingyangJiang, Yanchun Liang, Xiaoyue Feng, Xiaojing Fan, Zhili Pei, Yu Xue, and Renchu Guan; "Text Classification Based on Deep Belief and Softmax Regression"; RECENT ADVANCES IN PATTERN RECOGNITION AND ARTIFICIAL INTELLIGENCE; Springer, 2018.

[11]. ShufengHao ,Chongyang Shi , ZhendongNiu , Longbing Cao; "Concept coupling learning for improving concept lattice-based documentretrieval"; Engineering Applications of Artificial Intelligence 69 (2018) 65–75.

[12]. Sounak Banerjee, Prasenjit Majumder, and Mandar Mitra; "Re-evaluating the need for Modelling Term-Dependence in Text Classification Problems"; arXiv:1710.09085v1 [cs.IR] 25 Oct 2017.

[13]. Tom Kenter, Alexey Borisov and Christophe Van Gysel; "Neural Networks for Information Retrieval"; WSDM'18, February 5-9, 2018, Marina Del Rey, CA, USA.

[14]. Tsungnan Lin, Bill G. Horne, Peter Tiiio, and C. Lee Giles; "Learning Long-Term Dependencies

in NARX Recurrent Neural Networks"; JEEE TRANSACTIONS ON NElJRAL NETWORKS; VOL. I ,NO. 6, NOVEMBER 1996.

[15]. Vivi Nastase, Jelber Sayyad Shirabad and Maria Fernanda Caropreso; "Using Dependency Relations for Text Classi_cation";

[16]. Yemin Shi, Yonghong Tian, Yaowei Wang, Wei Zeng and Tiejun Huang; "Learning long-term dependencies for action recognition with a biologically-inspired deep network"; IEEE.

[17]. Yogesh Gupta, Ashish Saini , A.K. Saxena; "A new fuzzy logic based ranking function for efficient InformationRetrieval system"; Expert Systems with Applications 42 (2015) 1223–1234.

[18]. YoshuaBengio, Patrice Simard and Paolo Frasconi; "learning long-term dependencies with gradient descent is difficult"; IEEE; 1994.

Doaamabrouk" Enhancing Dependency using Power Set Theory for Text Classification" International Journal of Engineering Research and Applications (IJERA), Vol. 09, No.05, 2019, pp. 33-39