ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Efficient Document Classification using Phrases Generated by Semi Supervised Hierarchical Latent Dirichlet Allocation

[1] Rohit Agrawal, [2] A.S. Jalal, [3] S.C. Agarwal, [4] Himanshu Sharma
[1][2][3][4] GLA University, Mathura

***Abstract* - There are many models available for document classification like Support vector machine, neural networks and Naive Bayes classifier. These models are based on the Bag of words model. Word's semantic meaning is not contained by such models. Meanings of the words are better represented by their occurrences and proximity of words in particular document. So, to maintain the proximity of the words, we use a "Bag of Phrases" model. Bag of phrase model is capable to differentiate the power of phrases for document classification. We proposed a novel method to separate phrases from the corpus utilizing the outstanding theme show, Semi-Supervised Hierarchical Latent Dirichlet Allocation (SSHLDA).SSHLDA integrates the phrases in vector space model for document classification. Experiment represents an efficient performance of classifiers with this Bag of Phrases model. The experimental results also show that SSHLDA is better than other related representation models.**

Index Terms—Text classification, Latent Dirichlet Allocation, Semi Supervised Hierarchical Latent Dirichlet Allocation, Bag of word model, Bag of phrase model.

I. INTRODUCTION

Document Classification is able to automatically sort a set of documents into classes or categories. It is a supervised learning assignment i.e. useful for transmission labels to documents. Many methods are available for document classification which produces satisfactory result. Support Vector Machine (SVM) [1], Nearest Neighbor Classifier [2], Rocchio [3] and Naive Bayes Classifier [4] are popular classifiers which are used for document classification. According to past researches, it is observed that performance of SVM is better than other classifiers. Few improvements may still be developed in this classification. Bag of words model is used by all the above classifier for representing the text documents. Document is the basic unit for classification of text. This model contains the unordered collection of words. In the vector space model a global dictionary is used to represent documents where the number of words in the dictionary is represented by dimension of the vectors. It provides the efficient way for the representation of a document but the discriminative power of semantic meaning of two or more words that form phrases is ignored in this model. Likewise, these are not interpretable and far reaching. To catch discriminative power of words as expressions we require a model which can incorporate such n-grams in the vector space show with no extra changes in classifiers in light of the vector space models.

Ordered sequence of words is known as phrase. It is observable that if one word is combined with another word then it produces completely different meaning. For example, the word "stream" if appears as "data stream", the document is related to sequence of data packets that used to send or receive information. If stream appears as "river stream" then it is talking about large natural stream, may be waterway. Similarly "data mining" and "gold mining", "human race" and "bull race" and "nuclear reactor" and "nuclear bomb" are few sorts of expressions that containing a some basic word yet demonstrates an entirely unexpected importance when joined with different words. Latent Dirichlet allocation (LDA) [5] is able to find phrases from the document. But as we know datasets often grow over the time and when it grows they bring new entities and new structure so LDA is too rigid in this regard. LDA is an unsupervised model; it cannot take any information from hierarchical labels. Therefore, to remove this problem, we used a novel way to deal with discover phrases from the documents utilizing a subject model, Semi Supervised Hierarchical Latent Dirichlet Allocation [6].

It captures the breadth of useful topics across the corpus with the objective of organization of topics in the form of hierarchy. Semi-supervised hierarchical topic model, i.e. Semi-Supervised Hierarchical Latent Dirichlet Allocation (SSHLDA), is able to explore and find the latent topics from the documents and also take the information from the hierarchy labels to build the corresponding topics. We utilized these expressions together with singular words for Vector Space Model portrayal of the archives which is abused in order. Analyses demonstrate that Bag of Phrases show with proposed procedure beats the traditional Bag of

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

expression display. The paper is ordered as follow: Section 2 represents an overview of previous work in the field of document classification. Section 3 review the Bag of Words model, SSHLDA model and Bag of Phrases model. In the proposed methodology, we converse the algorithm based on SSHLDA to extract three word phrases in the corpus and how we can use it in classification, presented in section 4.The performance of proposed "Bag of Phrases" model using SSHLDA and represents the accuracy of SSHLDA for different classifier shown in section5. Finally, the last Section presents conclusions and future work.

2. RELATED WORKS:

There are many models accessible for characterization of documents however they all are utilizing the hidden "Bag of Words" display which overlooks the discriminative power of words in blend. Toward using discriminative power of mix of words Zhang et al. [7] displayed the idea of multi-word for characterization rather than singular words. They connected an algorithmic way to deal with extricate multi-words utilizing lexical instruments and accomplished a worthy exactness for characterization. Gunjariya et al.[8] had given a novel and basic plan for separating phrases utilizing theme models. Bag of Phrases Model catches the discriminative power of two words as a requested combine, but it is not able to capture the relationship between super topics and sub topics. In an unsupervised model it cannot take any information from hierarchical labels. There are many variations of topic models are available. They can be separated into four classes, for example, unsupervised hierarchical topic models, unsupervised non-hierarchical topic models and their comparing supervised hierarchical topic models and supervised non-hierarchical topic models as appeared in figure 1.

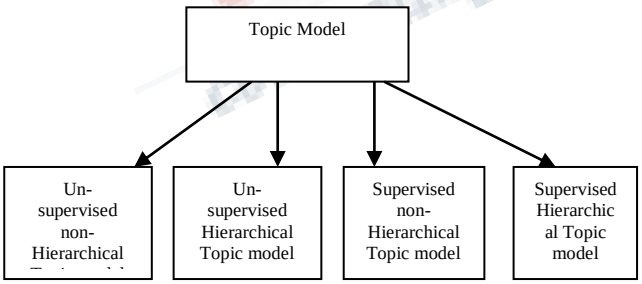

Figure 1: Classification of Topic Model

Unsupervised non-hierarchical topic models are LSA [9], pLSA [10], LDA [5], Hierarchical-concept TM [11] and Correlated TM [12] etc. The most famous is Latent Dirichlet Allocation (LDA). LDA is like pLSA, yet in LDA it is expected that the subject circulation have a Dirichlet earlier. LDA is a totally unsupervised calculation that models each archive as a blend of subjects. Other famous model is Correlated Topic Model (CTM) that maintains the correlation between the topics and also learns them. But topics in CTM are not independent and covariance matrix parameters raise as the square of the number of topics.

Several modifications have been done in LDA. Two such models, Supervised LDA [13] and DiscLDA [14] are able to model documents associate with only a single label. Another modification to eliminate this problem was proposed by TM [15] and Partially LDA (PLDA) [16]. However, these models obtain topics directly with the labels so this problem is removed by Labeled LDA (LLDA) [17].

Be that as it may, all the above models are not ready to catch the connection between the super and sub topics (Parent and kid). To tackle this issue, many models have been projected to keep up the relationship, for example, Hierarchical LDA (HLDA) [18], Pachinko Allocation Model (PAM) [19], and Hierarchical PAM (HPAM) [20] and so on. They all keep up the relationship as a pecking order, for example, Directed Acyclic Graph (DAG) or the tree. Blei et al. proposed the hLDA show [21], it take in the structure of a theme progressive system and furthermore the subjects that are contained in chain of importance. So this calculation can be utilized to extricate subject chains of importance from substantial datasets. Albeit unsupervised theme models show the different points per report, yet they are not suitable to consolidate the watched marks into their learning strategy. In this way for keep up the named data hLLDA [21] proposed. hLLDA can't catch the relations amongst parent and kid hub utilizing parameters and it not ready to recognize inactive subjects in the information space. To evacuate the disadvantage of hLDA and hLLDA and for take its advantage, Mao et al. proposed SSHLDA model [6], it can include the named subjects into the generative procedure. Then again, as hLDA, SSHLDA can naturally investigate inert theme in information space, and it can expand the current progression of watched subjects. So we used SSHLDA for express extraction this is a novel and direct arrangement for extracting phrases. The projected Bag of Phrases Model gets the discriminative vitality of three words.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

3. PRELIMINARIES:

3.1. Bag of word model:

Bag of words model is used by all the above classifier for representing the text documents. Usually, bag of word model represent by text in the vector space model which is represented by individual words that are obtained from the given text data set. As a simple and intuitive method, Bag of word method makes representation and learning highly efficient and easy because it ignores the order of the individual words. But this bag of word model is not efficient to represent the particular meaning and semantics which are available in documents. So it is necessary to remove this problem by using multiword in place of individual words. As we know semantic is better represent by the proximity of the words. "Document" is the basic unit for classification of text. Bag of words models contain the unordered accumulation of words. Vector space demonstrates is a mathematical model that represents any content archives as vectors of identifiers i.e. index terms. It is utilized as a part of many fields like ordering and importance rankings, data recovery and data sifting.

The Bag of Words model Provide the efficient way for the representation of document. But the discriminative power of semantic meaning of two or more words that form phrase is ignored in Bag of words models. When the document is scanned then dictionary of words is created.

3.2. The semi supervised hierarchical topic model:

Hierarchical topic modeling is able to maintain the relationship between parent-child and sibling topics. Unsupervised hierarchical topic modeling detects the automatically latent topic from the corpus. Hierarchical latent dirichlet allocation (hLDA) automatically obtains the hierarchy of topics. Unsupervised hierarchical topic modeling cannot take any information from the hierarchical labels. To remove this problem, supervised hierarchical topic modeling is used, named hierarchical labeled latent dirichlet allocation (hLLDA). HLLDA use hierarchical label to obtain the topics for each label but it is not able to find the latent topic in the data space.

As we know that from the full iceberg only some part of the iceberg is visible and rest part of the iceberg will be unseen. So it is necessary to identify or relate the document to all topics those are present in document. Semi-Supervised Hierarchical Latent Dirichlet Allocation (SSHLDA) is a probabilistic graphical model that portrays a procedure for creating a hierarchical labeled document collection. Like hierarchical Labeled LDA (hLLDA) [21], SSHLDA used Gibbs sampler [22] to provide the method for exploring the latent topics of the text corpus. It is based on Markov chain Monte Carlo algorithm [23] and obtains the sequence of observations when direct sampling is not possible. It is randomized algorithm and it generates random numbers so it produces different result each time. It produces the best value of the parameters, such as determining the number of patients admitted in particular hospital on a particular day. Gibbs sampler is divided into two steps- first is sampled the label allocation and second is sampling the path assignment. SSHLDA, first sample the path for particular document and identifies the word distribution of the topics for that path system. SSHLDA can add the labeled topics into the generative process. Then again, as various leveled Latent Dirichlet Allocation (hLDA) [18], SSHLDA can naturally investigate latent point in information space, and it can expand the current hierarchy system of observed topics. So SSHLDA can utilize the two kinds of subjects observed and latent.

3.3. Bag of phrase model and classification:

Bag of word model does not capture the discriminative power. We cannot extract the efficient information from the document. It neglects the semantic meaning of the particular word. Discriminative power of the word is dependent on the proximity of the words. When two or more words will be near to each other then they are able to show the efficient meaning rather than when used individually. Text classification is done by capturing the semantic difference between the words.

4. THE PROPOSED FRAMEWORK:

Our anticipated Bag of Phrases Model utilizing SSHLDA catches the discriminative control of three words as a requested combine and it ready to extricate covered up and also observed topics from the dataset. The proposed framework for extract the phrases from the document are shown in Figure 2. This framework consists of four phases: In first phase, we select the document or the text corpus then apply the SSHLDA topic model on it. SSHLDA provides the topic matrix, it contain both types of topics (observed and latent). The second phase is applying the phrase extraction (PE) algorithm on topic matrix and extracts the relevant phrases, here we use three word phrases extraction algorithm by using SSHLDA. In third phase, we integrate all retrieve phrases in vector space model so append this phrase list *g* in the old dictionary. New vector space model will be a dimension of (d+g) that represents the Bag of phrase model and document will be represented in the form of vector (r_1,r_2,r_3…..r_d, g_1,g_2,g_3….g_d) then standard classifier is applied

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

in the last phase. There are many classifiers are available for document classification like Support vector machine, neural networks etc. used to retrieve the relevant document from the large text collection or the text corpus. Unsupervised technique is useful to identify the hidden topics from the text corpus but it is not able to identify the information from observed label. But, supervised technique is able to take information form observed label. Therefore; we utilized the benefit of both techniques and applied Semi Supervised Hierarchical Latent Dirichlet Allocation topic model (SSHLDA) which is the combination of supervised and unsupervised learning.

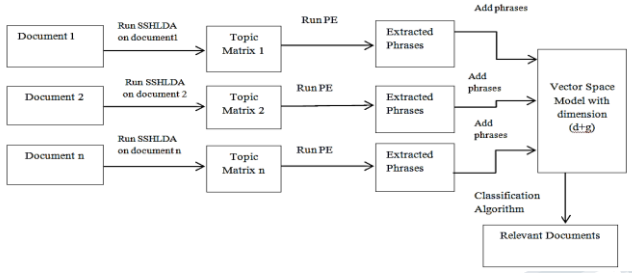

Figure 2: Framework for generates the phrases

The projected model, SSHLDA catches the discriminative power of three words as an ordered pair. Example is shown in figure 3 that dataset contains the documents about some topics and hidden words. Here, Observed topics are represented by gray squares and hidden topics represented by white squares. By using SSHLDA, we can generate the hierarchical labeled documents. Mohandas Karamchand Gandhi is a three words phrase extracted from text corpus. This is an observed topic from the dataset and its related latent topics are Bapu, Rashtrapita and Satyagrah Andolan. These all latent topics are generated automatically by using SSHLA model and also related to Mahatma Gandhi. So by using SSHLDA, we can identify the observed topics and latent topics in hierarchical form from the text corpus.

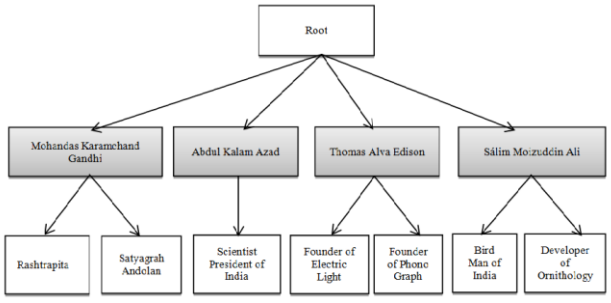

Figure 3: Topic distribution in hierarchical form

First create a word pair list then keep all words-pair that occurs together and find phrases. We insert all phrases in the tuples, phrases will be represents in the form of $<r_1,r_2,r_3>$, where r_1 is the first word and r_2 is the second word and r_3 is the third word of the phrase. In dataset, it is conceivable that a specific word is trailed by more than single word in a tuple. So it is possible that first word may be followed by different words in different tuples.

For extract the phrase first run SSHLDA on the given dataset and generates the topic matrix Φ. This topic matrix gives a word-topic distribution. In this topic matrix Φ each column contains a distribution of words for particular topic. Table 1 shows the symbols and their description used throughout of this paper.

Table 1: Notations and Their Description

Symbol	Description
D	Collection of documents
m	Combination of labels and words in document m
r_m	It represents the text of the particular document
s_m	Collection of topics in document m
δ_{si}	Multinomial distribution of the s_{i-1} sub topics
ψ_{si}	Dirichlet prior of δ_{si}
Θ	Distribution of words for u
ρ	Dirichlet prior of β
Φ	It represents the topic matrix
β	Multinomial distribution of words
$u_{m,n}$	For L available topics assign the n^{th} word in the m^{th} document
s_i	i^{th} level topic in hierarchy
U	All documents have a set of $u_{m,n}$ for all words
τ	Threshold
γ	It is a positive scalar

Here number of phrases depends on the threshold. If size of threshold will be high than number of phrases will be less or number of phrases will be more, if size of threshold is low. Here quality of phrases will be depends on the size of the threshold. Correspond threshold control the number of phrases in g and also paying attention on the excellence of phrases. Value of threshold will be set according to the performance of classifier. Following proposed algorithm is used for generating the phrase using SSHLDA.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Algorithm 1: Phrase extraction (PE) using SSHLDA

1. Create the word topic matrix Φ of dimension d*k, make tuple in the form of <r_1,r_2, r_3>
   (a) Select topic using $\beta_k \sim Dir(\rho)$
2. For i=1 to j=k do
   Arrange all columns Φ_j then make the inventory of crown words t
   End for
3. From all documents, m ∈{1, 2, 3,……,D}
   (a) Select s_1 as the root node.
   (b) All other labels will be denoted as l ∈ {2, ...,L}
      (i) Create a node s_l from Mult($\delta_{sl-1}|\psi_{sl-1}$), if node has been observed in particular level.
      (ii) Or draw a table s_l using

$$p(s_d = k|s_{1:(d-1)}) \propto \begin{cases} m_f & if\ f\ is\ occupied \\ \gamma & if\ f\ is\ new \end{cases}$$

4. Create a list of topics T.
   (a) Using Mult(Θ), draw U∈ {1,2,….,L)
   (b) Using topic, draw r_n.
5. For each word r_i∈list l_j do
   (a) $r_1 = r_i$
   (b) If $P(r_2| U = j) \geq \tau$ where r_2 is correspond to r_i
   (c) Add these phrase to list h
   (d) if $p(r_3|U=j) \geq \tau$ where r_3 is correspond to list h
   (e) Add these phrases to list g

Then compare it with threshold. If selected phrases are greater than to predetermine threshold ζ, and then add<r_1, r_2> to our list h. Otherwise discard it. Finally, compare the word r_3 to the list h's word. Check whether r_3 is corresponding word of h list or not from the topic matrix.

This algorithm is used for extract the three word phrase from the text corpus. All topics will be store in the form of hierarchy. Root node will be denoted by s1 then labeled all other documents. Arrange all topics in hierarchical form. For any document top related 10-15 words of topics will be good descriptor in hierarchy. There Root and Mohandas Karamchand Gandhi has a parent child relationship between their topics. There Bapu, Rashtrapita are the words which are more related to topic Gandh ji. These all relationship is maintained in the form of hierarchy. We have used the SVM classifier for classification of documents. It is observed that performance of SVM is better than other classifier.

5. EXPERIMENTAL RESULTS AND ANALYSIS:

5.1 Data Preprocessing

The proposed approach is tested on 50 documents' dataset. As per the best knowledge to till date, no manually segmented dataset for three word phrases is publicly available. So to obtain the comparative results for three word phrases using different dataset is difficult. As the ground truth of this dataset is not available so the results cannot be verified by comparing the results of various methods. We can verify the results by testing the existing methods on the same dataset which we have used. From given evaluation and comparison analysis, we can see that proposed approach provides better result than other approaches. The calculation complexity of the proposed approach is also easier and accurate than previous approaches.

These documents contain the information about the famous persons and there innovations. These documents are stored in text format. We have used the SSHLDA to extract the three word phrases from the dataset. Three word phrases provide more semantic meaning as compared to one word.

The proposed approach gives better results because it can detect the latent as well as observed phrases. To demonstrate that phrase extraction using SSHLDA, it is better to compare with the previous approaches because it can extract the latent as well as observed phrases and maintain the hierarchical relationship between all phrases; there are some phrases to show the results. There are some samples of the phrases (Observed as well as latent) are given in table 2.

Table 2: Some Extracted Phrases from Our Dataset

Phrases (Observed Topic)	**Phrases (Latent Topic)**	**Phrases (Latent Topic)**
Salim Moizuddin Ali	Bharatpur Bird Sanctuary	Bird Man of India
Satyendra Nath Bose	Bengali Indian Physicist	Developer of Quantum Mechanism
Thomas Alva Edison	Founder of Phono Graph	Founder of Electric Light
Ratan Naval Tata	Chairman of Tata Group	Chairman of Tata Sons
Dabbala Rajgopal Reddy	Turing Award Wining	Robotics Institute CMU
Sunil Bharti Mittal	CEO of Bharti Enterprises	International Chamber of Commerce
Homi Jehangir Bhabha	Father of Nuclear Programme	Atomic Energy Research

We need to mine phrases from the documents but we didn't stem the words. We find phrases using training data than used these phrases for text classification. Some examples of phrases are shown in table 2, these phrases are obtained by our phrase extraction algorithm and it is proved that the semantics are better captured by phrases in place of individual words. This approach extracts latent and observed topics and maintains the relation between them. The results

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

are explained with illustrative examples. The experimental results of existing approaches of phrase extraction and proposed approach are reprented in the form of tables.

5.2 Results and evaluation:

As mentioned above we used the SSHLDA with Support Vector Machine (SVM) because it works efficiently as compared to the other classifiers. For our experiment we used libsvm [24] with a linear kernel. These observations are used for large data sets. It is observed that if we increase the threshold then it improves the accuracy because all useless phrases will be pruned. But for particular level if threshold will be high than it will discard some important phrases and will get lesser accuracy in document classification.

In all experiments we are using 50 different documents i.e. k = 50 and number of words per document is 50 i.e. t=50. Table 3 shows that some classes are better performed for Bag of Phrase model in comparison to bag of word model. For phrase extraction of documents, we have used accuracy of phrases for the evaluation of the results. Accuracy of phrases shows the quality of the extracted phrases from the document. There are some phrases which are extracted from the dataset. These phrases are the combination of the latent as well as observed phrases and maintain the relationship between all phrases. So its accuracy for getting the correct document will be higher in comparison with the other approaches. We also calculated the value of precision and recall for same phrases. Then compare these phrases and check the accuracy, precision and recall of the phrases and at last compare the performance of our approach (BOP using SSHLDA) to previous approaches (BOW and BOP using LDA). Accuracy, Precision and Recall are calculated using following formulas:

$$\text{Accuracy} = (\text{Measured Topics} * 100)/(\text{Expected Topics}) \quad (2)$$

$$\text{Precision} = \#\text{ of TP}/(\#\text{ of TP} + \#\text{of FP}) \quad (3)$$

$$\text{Recall} = \#\text{ of TP}/(\#\text{ of TP} + \#\text{of FN}) \quad (4)$$

Where:

True Positives (TP) - Number of correctly identified documents.

True Negatives (TN)- Number of incorrectly identified documents.

False Positives (FP) - Number of correctly rejected documents.

False Negatives (FN)- Number of Incorrectly rejected documents.

Table 3: Accuracy of Extracted Phrases and comparison with other approaches

DataSet	Classifier	BOW (%)	BOP (LDA)	BOP (SSHLDA
Satyendra Nath Bose	SVM	69.56	73.25	80.00
Lakshmi Niwas	SVM	74.28	86.02	94.34
Dabbala Rajagopal	SVM	79.18	87.23	92.23
Sunil Bharti Mittal	SVM	87.24	91.12	95.46
Homi Jehangir	SVM	89.39	93.54	98.02

Table 4: Precision of Extracted Phrases and comparison with other approaches

DataSet	Classifier	BOW (%)	BOP (LDA)	BOP (SSHLDA)
Satyendra Nath Bose	SVM	74.56	85.25	97.69
Lakshmi Niwas Mittal	SVM	78.28	80.02	82.34
Dabbala Rajagopal	SVM	72.35	81.22	91.23
Sunil Bharti Mittal	SVM	76.24	85.12	94.46
Homi Jehangir Bhabha	SVM	81.39	88.54	95.02

Table 5: Recall of extracted phrases and comparison with other approaches

DataSet	Classifier	BOW (%)	BOP (LDA)	BOP (SSHLDA)
Satyendra Nath Bose	SVM	72.56	63.23	50.69
Lakshmi Niwas Mittal	SVM	68.28	66.59	62.34
Dabbala Rajagopal	SVM	74.23	65.12	59.65
Sunil Bharti Mittal	SVM	70.12	62.24	55.23
Homi Jehangir Bhabha	SVM	65.39	58.57	52.02

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

We have calculated two parameters: recall and precision. Precision is the probability that a retrieved document is significant and recall is the probability that relevant document is retrieved in search. The proposed approach has achieved 56.00% recall and 92.15% precision. For performance comparison, the results are compared with previous approaches.

Table 6: Comparison of Accuracy, Precision and Recall with previous approaches

Approaches	Accuracy (%)	Precision (%)	Recall (%)
BOW	79.93	76.56	70.12
BOP (LDA)	86.23	84.03	63.15
BOP (SSHLDA)	92.01	92.15	56.00

This approach can also detect the latent as well as observed phrases and maintain the relationship between all phrases. So its accuracy is higher comparison the other approaches. Then results are better in the proposed approach.

6. CONCLUSIONS AND FUTURE WORK:

We have utilized a basic and proficient plan for extracting phrases. The proposed Bag of Phrase model can extract the phrases and demonstrates the discriminative power of the words as a requested combine. It is necessary to use Bag of Phrase model because Bag of Word model fails to shows the discriminative power of the words and semantic effect of the words, when used in combination. This topic model i.e. Semi Supervised Hierarchical Latent Dirichlet allocation is able to remove the drawback of hLDA and hLLDA and combine their advantages. SSHLDA is able to shows the information of labels and also explore the latent topics from the corpus. This provides the efficient result for classifiers like Support Vector Machine and Nave Bayes Classifier and classification accuracy is also improved here. For improve the effectiveness, our next target is to extend the length of the phrases and explore new topic models for hierarchical data.

REFERENCES

[1] C. J. Burges, “A tutorial on support vector machines for pattern recognition”, Data Mining and Knowledge Discovery, 2:121–167, 1998.

[2] B. Dasarathy, “Nearest neighbor (fNNg) norms:fNNgpattern classification techniques” 1991.

[3] T. Joachims, “A probabilistic analysis of the rocchio algorithm with tfidf for text categorization”,Journel of Machine Learning Research, 14:143–151, 1997.

[4] I. Androutsopoulos, J. Koutsias, and Chandrinos, “An evaluation of naive bayesian anti-spam filtering”,Arxiv preprint cs/0006013, 2000.

[5] D. M. Blei, A. Ng, and M. Jordan, “Latent dirichlet allocation”,.Journel of Machine Learning Research, 3:993– 1022, 2003.

[6] D. Wang, M. Thint, and A. Al-Rubaie, “Semi-Supervised Latent Dirichlet Allocation and its Application for Document Classification,” IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, IEEE, Vol. 03, pp. 306-310, 2012.

[7] W. Zhang, T. Yoshida, and X. Tang, “Text classification based on multi-word with support vector machine. Knowledge-Based Systems”, Elsevier, 21:879– 886, 2008.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[8] D.Gujraniya, and M.N.Murty, "Efficient classification using phrases generated by topic models." Proceedings of the 21st International Conference on Pattern Recognition, IEEE, pp. 2331-2334, 2012

[9] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, “ Indexing by latent semantic analysis”, Journal of the American society for information science, 41(6):391–407, 1990.

[10] T. Hofmann, “Probabilistic latent semantic analysis” In Proc. of Uncertainty in Artificial Intelligence, UAI’99, page 21. Citeseer,1999.

[11] C. Chemudugunta, P. Smyth, and M. Steyvers,”Text modeling using unsupervised topic models and concept hierarchies” Arxiv preprint arXiv:0808.0973, 2008.

[12] D. Blei and J. Lafferty, “Correlated topic models”, Advances in neural information processing systems, 18:147, 2006

[13] D.M. Blei and J.D. McAuliffe, "Supervised topic models", In Proceeding of the Neural Information Processing Systems(nips),2007.

[14] S. Lacoste-Julien, F. Sha, and M.I. Jordan,"ndisclda: Discriminative learning for dimensionality reduction and classification",Advances in Neural Information Processing Systems, 21,2008.

[15] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth," The author-topic model for authors and documents", In Proceedings of the 20th conference on Uncertainty in artificial intelligence, pages 487–494. AUAI Press,2004.

[16] D. Ramage, C.D. Manning, and S. Dumais," Partially labeled topic models for interpretable text mining", In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 457–465. ACM, 2011.

[17] D. Ramage, D. Hall, R. Nallapati, and C.D. Manning, "Labeledlda: A supervised topic model for credit attribution in multi-labeled corpora", In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, pages 248–256. Association for Computational Linguistics, 2009.

[18] D. Blei, T.L. Griffiths, M.I. Jordan, and J.B. Tenenbaum., "Hierarchical topic models and the nested chineserestaurant process", Advances in neural information processing systems, 16:106,2004.

[19] W. Li and A. McCallum, "Pachinko allocation: Dag-structured mixture models of topic correlations", In Proceedings of the 23rd international conference on Machine learning, pages 577–584. ACM, 2006.

[20] D. Mimno, W. Li, and A. McCallum," Mixtures of hierarchical topics with pachinko allocation", In Proceedings of the 24th international conference on Machine learning, pages 633–640. ACM,2007.

[21] Y. Petinot, K. McKeown, and K. Thadani, "A hierarchical model of web summaries", In Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies: short papers-Volume 2, pages 670–675. ACL, 2011.