International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-7 Issue-11, August 2018

An Efficient Approach to Reduce Text Dimension for Precise Text Classification for Big Data

Akshada Bhanushali, Pravin Rahate

Abstract: ***In today's society, few famous news websites such as Google and sina server gives information for users. But recently with the continuous development of information technology, the quantity of disorder data is increasing in volume. Text classification and organization has become a task. The traditional manual classification of news text not only consumes a lot of human and financial resources, but classification is also not achieved quickly. This paper makes a research about the news text classification. A news text classification model is proposed based on Latent Dirichlet Allocation (LDA) and Domain Word Filtering. The model reduces the features dimension of the news text effectively and gets good classification results. This model uses topic model to reduce text dimension and get good features as the dimension of the news texts is too high.***

Keywords: Topic Model, LDA, Domain Word Filtering, News Website, Text Classification

I. INTRODUCTION

Nowadays with the continuous development of information technology, there is an explosive increase in the information data of internet. The main platform for human to get news information is the major news websites. The news data of news portal is increasing, which also brings some challenges to the site are because; the news data of news portal is increasing. The traditional text classification methods are not able to meet the needs of the current social development. So the hot topic in the field of the text mining in recent years is research on text classification model [1]. The news text classification system can quickly handle the text data fast, and can also make accurate prediction of the classification labels. So in this way automatic classification can help to complete text classification function for news platform with excellent efficiency, and it can also help the company to save expenses. The research on automatic text classification plays an increasingly important role in the era of big data. Classic text classification algorithms which have been proposed and widely used are like support Vector Machine, k-Nearest Neighbor, Naive Bayes and Decision tree and etc. But each kind of classification algorithms has different strengths and weaknesses. Different classification algorithms for different scene are decided by its strengths and weaknesses. General steps of text categorization are shown in Fig.1:

Manuscript published on 30 August 2018.
*Correspondence Author(s)
Akshada Bhanushali, PG Student, Department of Computer Engineering, Datta Meghe College of Engineering, Navi Mumbai (Maharashtra), India. E-mail: aksha.bhanushali24@gmail.com
Pravin Rahate, Assistant Professor, Department of Computer Engineering, Datta Meghe College of Engineering, Navi Mumbai (Maharashtra), India. E-mail: psr.cm.dmce@gmail.com

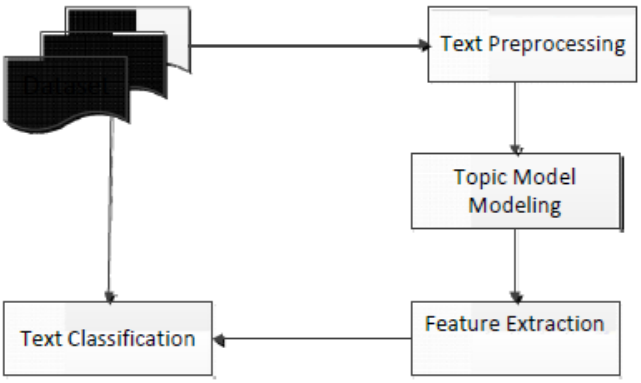


Fig. 1. Example of Classification Process

II. LITERATURE REVIEW

Latent Dirichlet Allocation (LDA) is a of topic model algorithm based on probability model [2]. The algorithm thinks that each article is composed of a plurality of topic mixture. It can catch the potential hidden information topic in large-scale document set. The algorithm assumes that each word in the mass of an article is through by "with a certain probability to choose a topic from the data, and then from this subject with a certain probability to select a word from the content". It chooses a topic from topic distribution and then chooses a word from the word distributions from corresponding topic [3].

The above procedure is repeated until it makes the traversal of the document for every word. θ represents topic distribution and φ represents word distribution.

The generative process of LDA is given as follows:

- *Choose* $\theta \sim Dir(\alpha)$
- *Choose* $\varphi \sim Dir(\beta)$

For each word in the text:

1. *Select a topic* $z \sim Multinomial(\theta)$
2. *Select a words distribution according to the topic z*
3. *Select a word* $w_n \sim Multinomial(\varphi_z)$
4. *Repeat it until each word of a text document*

The idea of the algorithm is to select a topic vector θ and the topics determine the probability of each subject which is selected. Then it can choose a topic z from θ. Finally at the end it chooses a word ω from the corresponding word distribution of corresponding topic distribution [4].

The process of LDA is summarized below:

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

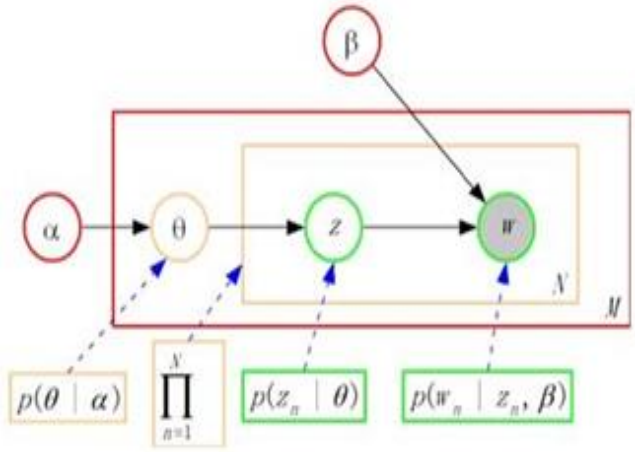


Fig. 2. Flow of Classification Process

The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document

M= number of documents
N= total number of words in all documents
α= super parameter of topic distribution
β= super parameter of word distribution
Z= identity of topic of all words in all documents
P(w | z)= Dirichlet distribution

LDA joint distribution formula is following:

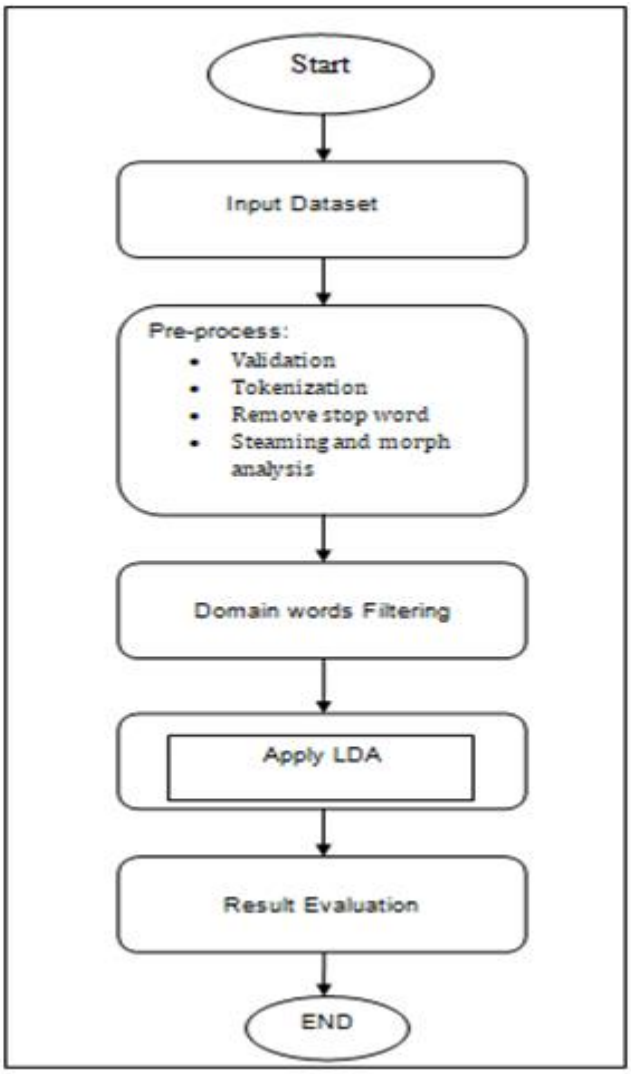
$$p(\theta, z, \omega \mid \alpha, \beta) = p(\theta \mid \alpha) \prod p(z_n \mid \theta) p(w_n \mid z_n, \beta)_{n=1}$$

Topic model is a unsupervised learning method. But text classification is mostly used in supervised learning. So in this paper we will use the supervised topic model, which is, Latent Dirichlet Allocation Supervised (SLDA) and the class labels are from experiment data [5]. But the Latent Dirichlet Allocation Supervised (SLDA) model uses continuous response values by linear regression which cannot be applied for multi-class text data like news text.

III. PROBLEM DEFINATION

The dissertation aims at implementation of System, which classify the documents based on its content. As the data on the Google news changes at regular intervals we need Scalability and Efficiency while classification of the documents, as there is huge problem in handling the tremendous data growth. In order to handle the above mentioned problem efficiently, we propose to maintain domain specific dictionary with the help of Domain Word Filtering. This dictionary will analyze article from selected domains and after pre-processing the documents and by calculating the term frequency, and with the help of dictionary clustering is done by using LDA algorithm. Our system gives the better results for the document classification for Big data.

IV. SYSTEM ARCHITECTURE



In the above model the initial step is input data, the dataset can be in any number of packets. The dataset covers a wide range of any topics like religion, medicine, hospital, arts and so on. Preprocessing step includes validation, tokenization, removing stop word, steaming and morph analysis. Validation includes elimination of non English words in the dataset, in which all the non English words will be removed.

Tokenization takes place after validation where token will be given to the words from the dataset using predefined java class. Also special characters and non English words like /, % will be removed. We use regular expression to remove special character. ASCII code will be used to remove all the non english words. For stop word removal there will be backend stop word dictionary will filter those and remaining words will be processed. Steaming and morph analysis will be done by snowball in which string processing language will be used. For example hopefulness will be hopeful similarly relational will be relate. After this Domain word filtering will be done. The brief information about domain word filtering will be below. Then LDA algorithm will be applied and finally will get the output.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

V. EXPERIMENTAL SETUP AND RESULT

To improve result of system we are adding this module. Domain word filtering helps us to filter unwanted words and it gives only those words for which system is trained. This process directly effect on system performance and result. The paper uses the precision, recall and F1-Measure as a measure of predictive performance evaluation index to examine assess algorithm for news text classification. Precision and recall are the basic measures used in evaluating search strategies.

$$STP(ci)recall = \sum_{ui \hat{I} U} \frac{ST(ci)}{ui \hat{I} U}$$

$$STP(ci)Precision = \sum_{ui \hat{I} U} \frac{SL(ci)}{ui \hat{I} U}$$

$$F1 = \frac{2' recall' precision}{Recall + precision}$$

$$= \frac{2 \sum TP(ci)}{\sum_{ui \hat{I} U} ui \hat{I} U}$$

$$\sum_{ui \hat{I} U} SR(ci) + \sum_{ui \hat{I} U} ST(ci)$$

$^{TP}(ci)$ Represent the number of correct prediction of class i

$^{L}(ci)$ Represents the return number of documents by class i by classification model.

$T(ci)$ Represents the real document number of class i

Recall is the ratio of the number of relevant record retrieved to the total number of relevant records in the database. Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. Recall and precision both are expressed in percentage. F1-Measure is the key index of the experimental result. F1-Measure is the harmonic mean of precision and recall.

Below table is the analysis in which first table gives result by using only LDA algorithm and second table gives result by using LDA algorithm and Domain word filtering by which the accuracy increases by 6% and gives minimum word count.

Table 1.Experimental Analysis with LDA Algorithm

Domain	Health	Entertainment	Sports	Technology
TN	28	30	29	30
TP	7	5	7	6
FN	4	3	1	2
FP	1	2	3	2
Precision	0.875	0.71428571	0.7	0.75
Recall	0.6363636364	0.625	0.875	0.75
P*R	0.5568181818	0.4464285714	0.6125	0.5625
P+R	1.511363636	1.339285714	1.575	1.5
Fmeasure	0.7368421053	0.6666666667	0.7777777778	0.75

As shown in the Table, the experimental data is from the Google news group data packets in complete text format which is used in classification algorithm. The dataset cover a wide range of mix topics. The dataset can be in any wide range like MB or GB. The model chooses different kinds of categories in our experimental dataset.

Table 2. Experimental Analysis with LDA Algorithm and DNF

Domain	Health	Entertainment	Sports	Technology
TN	28	30	29	30
TP	9	8	7	8
FN	2	0	1	0
FP	1	2	3	2
Precision	0.9	0.8	0.7	0.8
Recall	0.8181818182	1	0.875	1
P*R	0.7363636364	0.8	0.6125	0.8
P+R	1.718181818	1.8	1.575	1.8
Fmeasure	0.8571428571	0.8888888889	0.7777777778	0.8888888889

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The main focus is parameters of the algorithm also have influence on the parameter and the experimental dataset. At the same time the most important specification is the number of topics in complete data.

The paper focuses on the topic model and also extends LDA plus Domain Word Filtering to a multi-class supervised topic model. This can be helped to classify into multiple classes.

VI. CONCLUSION

We proposed a news text classification based on LDA topic model with Domain Word Filtering, which is mostly based on latent topic information of the text. The ultimate goal is to reduce text dimension for text classification.

The news text classification system can quickly handle with all text data fast, and make accurate prediction of the classification labels from large content. The topic model reduces the text vector from higher dimension to lower dimension and also chooses the good features attribute of the news text. Thus with the help of Domain Word Filtering accuracy is more also time is saved. Automatic classification can help to complete text classification function for news platform with high efficiency, and it can also help the company to save expenses also it is very useful in all ways to reduce the features dimension of the news text and get good classification results, to produce good quality clusters and improves the scalability and efficiency.

In this project initially we had applied for small and one dataset in which approximately 42% words got reduced after applying Domain Word Filtering and now this can be applied for Big Data for News, sports and Medicine Domain in which recall, precision and F1 measure will be predicted. Also we are working to use this in various other domains like in Entertainment, Education for better accuracy in future scope.

REFERENCES

1. Zhenzhong Li “News text classification model based on topic model” in Communication University of china Beijing, china, 2016.
2. S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, “Data mining techniques and applications–a decade review from 2000 to 2011,” Expert Systems with Applications, 2012
3. G. Kumaran and J. Allan, “Text classification and named entities for new event detection,” in SIGIR, 2004.
4. X. Liu, B. Huet “Heterogeneous features and model selection for event- based media classification,” in ICMR, 2013.
5. J. R. Kender and M. R. Naphade, “Visual concepts for news story tracking: Analyzing and exploiting the nist trecvid video annotation experiment.” in CVPR, 2005.
6. D.M. Blei, A.Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” JMLR, vol.3, pp. 993-1022, Mar. 2003
7. S. Wenqian, D. Hongbin, Z. Haibin, and W. Yongbin, “A novel feature weight algorithm for text categorization,” in Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE ’08), pp. 1–7, Beijing, China, October 2008.
8. G. King, P. Lam, and M. Roberts, “Computer-assisted keyword and document set discovery from unstructured text” (2014)
9. N. Padhy, D. Mishra, R. Panigrahi et al., “The survey of data mining applications and feature scope,” (2012)

Akshada Bhanushali Pursuing ME from Datta Meghe College of Engineering, in Computer Engineering & had completed BE from Lokmanya Tilak college of Engineering (Computer Science & Engineering) in the year 2015, & interested in research Text Classification. Our first paper got published in International Conference on Recent Development in Computer and Information Technology (ICRDCIT) on 03rd April 2018 Mumbai, India.

Prof Pravin Rahate currently is the Professor at Datta Meghe College of Engineering for Computer science department and is the Project guide for this topic. Our first paper got published in International Conference on Recent Development in Computer and Information Technology (ICRDCIT) on 03rd April 2018 Mumbai, India.