# SNA: Construction of Meta classifications in Domain Modeling

P.Sasikala[1], Dr.P.Mayilvahanan[2]

Research Scholar, Vels Institute of Science, Technology & Advanced Studies,
VISTAS, Pallavaram, Chennai[1]
Professor, Department of Computer Science, Vels Institute of Science,
Technology & Advanced Studies, VISTAS, Pallavaram, Chennai[2]

**Abstract - By using text mining concepts to be found out the domain modeling depended on the vocabularies and relevant domains from the documents. Each document is having various concepts and each concept is having various terminologies. Collection of wide documents are very useful and strongly to b0e understood in that particular domain. Sometimes documents are associated with meaningful or meaningless information in the particular domain. Machine learning one of the interested concepts is domain modeling. This research work focuses on the huge dataset for author article relationship with appropriate Ensemble classification approaches applied and find out the best accuracy of this dataset.**

**Keywords:** Data Mining, Meta, Domain Modeling.

## I. INTRODUCTION

Topic modeling helps in exploring large amounts of text data, finding clusters of words, similarity between documents, and discovering abstract topics. As if these reasons weren't compelling enough, topic modeling is also used in search engines wherein the search string is matched with the results. A Topic Model can be defined as an unsupervised technique to discover topics across various text documents. These topics are abstract in nature, i.e., words which are related to each other form a topic. Similarly, there can be multiple topics in an individual document.

This black box (topic model) forms clusters of similar and related words which are called topics. These topics have a certain distribution in a document, and every topic is defined by the proportion of different words it contains.

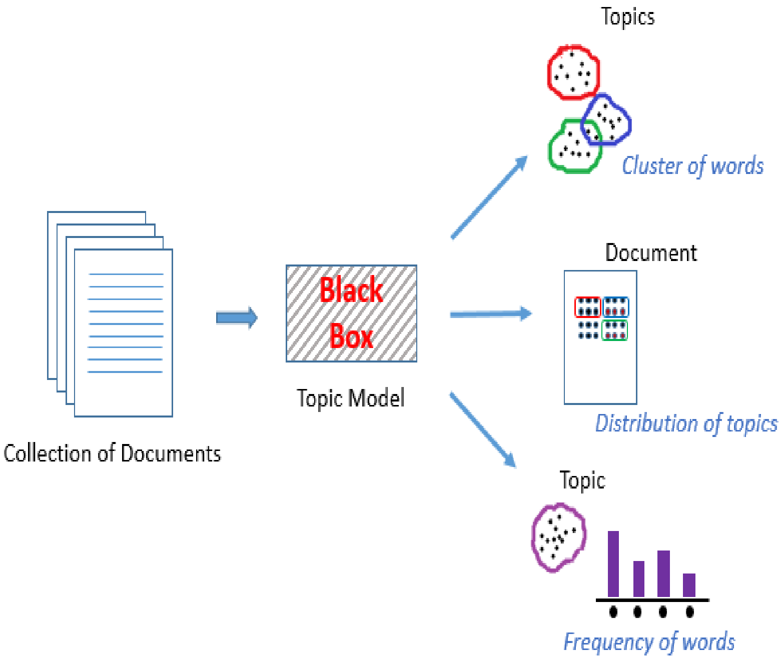

Figure 1: Topic Model

Figure 1 shows the increasing trend of cross-domain collaborations over the past fifteen years across different domains in a publication database In most of the cases, there exists a clear increasing trend of the cross-domain collaborations. The present study was focused on understanding the author's collaboration among research community in Aminer dataset.

The rest of this paper is organized as follows: Section 2 represents the materials and methods; Section 3 presents our results and discussions; then conclusion presents in Section 4.

## II. MATERIALS AND METHODS

In this section presents the materials and methods of this research work. Two components were considered in this section.

1. Extracting the themes in the article's abstract by domain modeling (this experiment was realized by using the tool **MeSH (Medical Subject Headings) (Domain Extraction Tool)**)

2. Classifying by using weka 3.8.3 version those identified domains into the domain subject area.

**Dataset Information:**

The dataset collected was named as Topic__paper_author in the academic social network data from https://aminer.org/domain_paper_author shown in Table 1.

Table 1 Description of Topic_Paper_Author Dataset

| S.No. | Attribute with Datatype |
|---|---|
| 1 | Conference Name(Text) |
| 2 | Title(Text) |
| 3 | Year(Numeric) |
| 4 | Abstract(Text) |
| 5 | Authors(Text) |

The dataset has collected for the purpose of cross domain recommendation depicted in table 2.

Table 2 Details of Domain with conferences and authors

| S.No | Domain | Conferences | Authors and Co authors |
|---|---|---|---|
| 1 | Data Mining | KDD, ICDM | 6,282 &22,862 |
| 2 | Medical Informatics | WSDM | 9,150&31,851 |
| 3 | Theory | PKDD, FOCS, SODA | 5,449&27,712 |
| 4 | Visualization | CVPR, ICCV, VAST, TVCG | 5,268&19,261 |
| 5 | Database | SIGMOD, VLDB, ICDE | 7,590&37,592 |

There are 10 classes in the domain attributes were shown in Table 3.

Table 3 Description of 10 classes in domain attributes

| S. No. | Class with No of records | 10 % |
|---|---|---|
| 1 | Algorithm(3999 records) | 399 records |
| 2 | Artificial Intelligence(630 records) | 63 records |
| 3 | Biomedical informatics(961 records) | 96 records |
| 4 | Clinical(224 records) | 22 records |
| 5 | Data Mining(2496 records) | 250 records |
| 6 | Database(4635 records) | 464 records |
| 7 | Image Processing(4064 records) | 406 records |
| 8 | Medical Imaging(1159 records) | 116 records |
| 9 | Programming(110 records) | 11 records |
| 10 | Telemedicine(97 records) | 10 records |

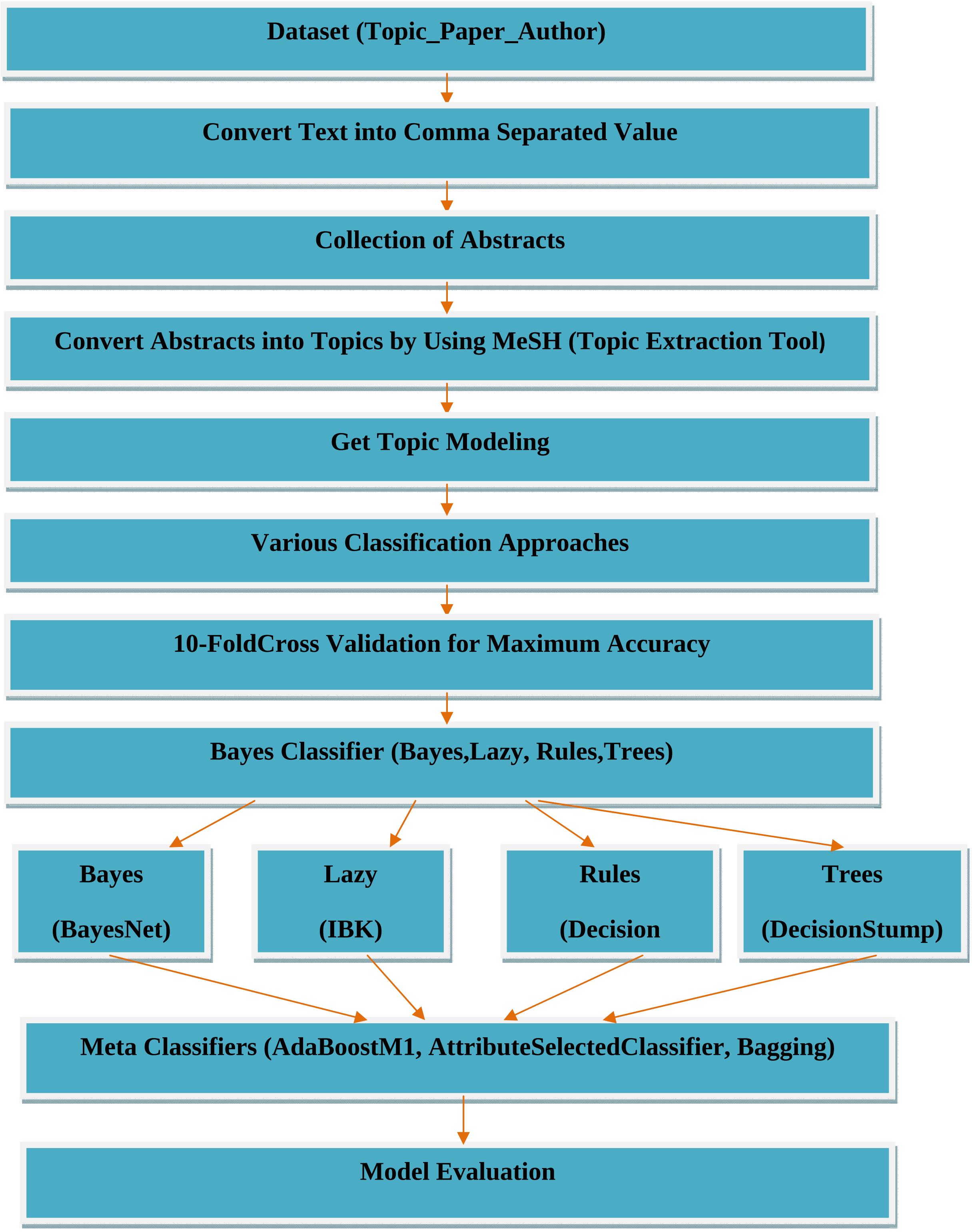
Dataset (Topic_Paper_Author)
Convert Text into Comma Separated Value
Collection of Abstracts
Convert Abstracts into Topics by Using MeSH (Topic Extraction Tool)
Get Topic Modeling
Various Classification Approaches
10-FoldCross Validation for Maximum Accuracy
Bayes Classifier (Bayes,Lazy, Rules,Trees)
Bayes (BayesNet)
Lazy (IBK)
Rules (Decision
Trees (DecisionStump)
Meta Classifiers (AdaBoostM1, AttributeSelectedClassifier, Bagging)
Model Evaluation

Figure 2 Proposed System Architecture

The second component in this experiment was comprehend by applying numerous classifiers for carry out the better classification accuracy for categorizing the subject catalog for the specified dataset, namely “Topic Paper Author” dataset. It has 18,375 instances with 5 attributes. In this research work 10% of records have taken from this dataset due to Weka 3.8.3 version heap memory limitations.

## III. RESULTS AND ANALYSIS

In this section discusses results and analysis of this research work. In AdaBoostM1 with BayesNet accuracy was 89.78%, AdaBoostM1 with IBK accuracy was 93.95%, AdaBoostM1 with DecisionTable accuracy was 94.55% and AdaBoostM1 with DecisionStump 44.64%. In AttributeSelectedClassifier with BayesNet accuracy was 94.50%, AttributeSelectedClassifier with IBK accuracy was 94.12%, AttributeSelectedClassifier with DecisionTable accuracy was 94.61% and AttributeSelectedClassifier with DecisionStump 44.64%. In Bagging with BayesNet accuracy was 94.39%, Bagging with IBK accuracy was 94.07%, Bagging with DecisionTable accuracy was 94.50% and Bagging with DecisionStump 94.50%.

Table 4 Ensemble Model Accuracies

| S.No | Meta Classifiers | Accuracy | Accuracy |
|---|---|---|---|
| 1 | AdaBoostM1 | BayesNet | 89.78% |
| 2 | | IBK | 93.95% |
| 3 | | DecisionTable | 94.55% |
| 4 | | DecisionStump | 44.64% |
| 5 | AttributeSelectedClassifier | BayesNet | 94.5% |
| 6 | | IBK | 94.12% |
| 7 | | DecisionTable | 94.61% |
| 8 | | DecisionStump | 44.64% |
| 9 | Bagging | BayesNet | 94.39% |
| 10 | | IBK | 94.07% |
| 11 | | DecisionTable | 94.5% |
| 12 | | DecisionStump | 94.5% |

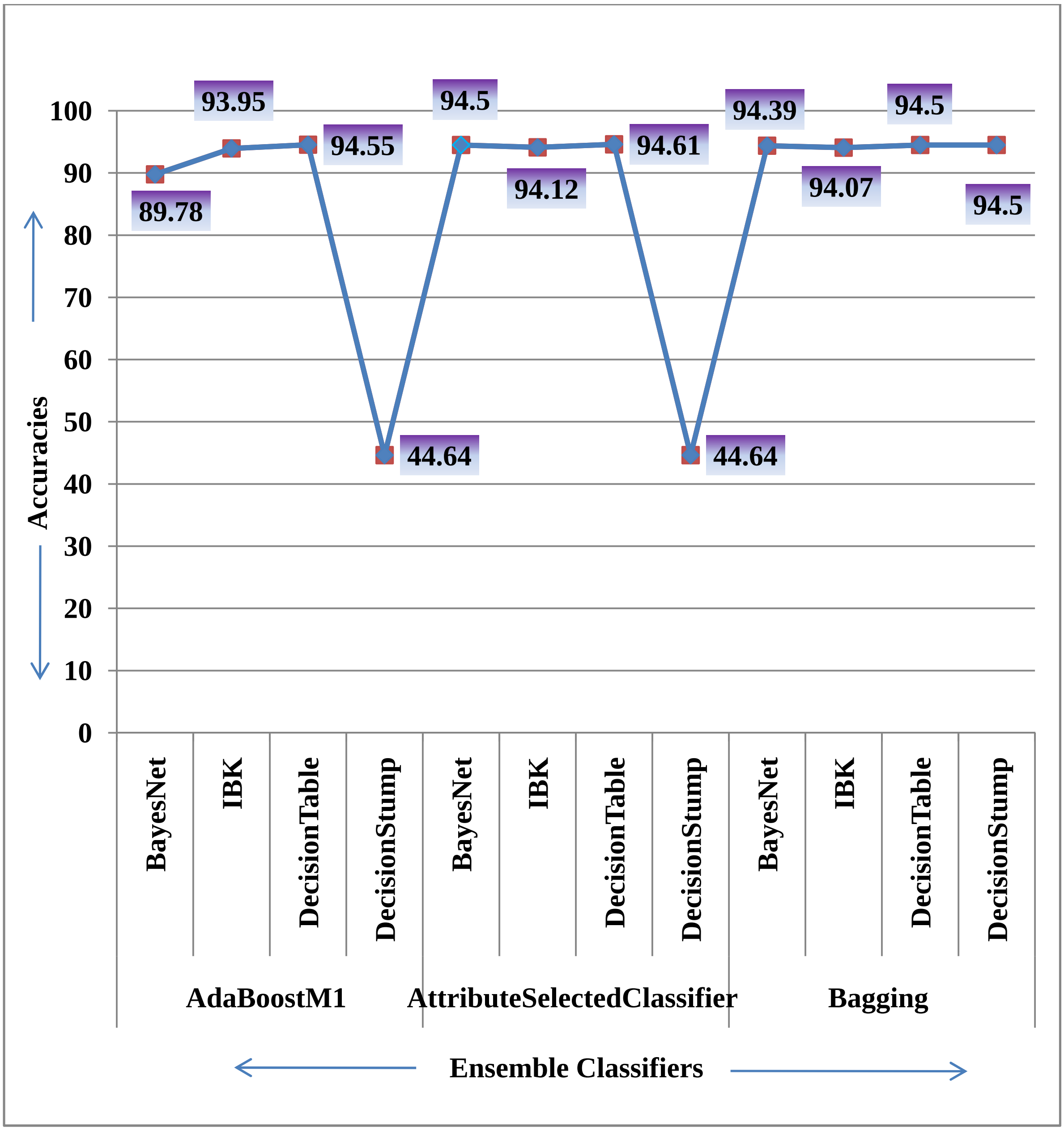

Figure 2 Graphical representations of various ensemble classifiers accuracy levels

The accuracies obtained from the selected classifiers are shown in Figure 2.This chart represents the comparison of all the categories of the classifiers. In AdaBoostM1 classifier with DecisionTable has high accuracy when compared with AdaBoostM1 with BayesNet Classifier model,AdaBoostM1 with IBK Classifier model and AdaBoostM1 with DecisionStump Classifier Model.

In AttributeSelectedClassifier with DecisionTable has high accuracy when compared with AttributeSelectedClassifier with BayesNet Classifier model, AttributeSelectedClassifier with IBK Classifier model and AttributeSelectedClassifier with DecisionStump Classifier Model.

In Bagging with DecisionTable and Bagging with DecisionStump have same as well as high accuracy when compared with Bagging with BayesNet Classifier model, and Bagging with IBK Classifier model.

Table 5 Ensemble Model Various Measures

| S.No | Name of the Classsifier | TP Rate | FP Rate | Precision | Recall | F Measure | MCC | ROC | PRC | Time Taken for Build for this Model (In Seconds) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | AdaBoostM1 with Bayes Net | 0.89 | 0.01 | 0.9 | 0.9 | 0.9 | 0.84 | 0.92 | 0.8 | 0.77 |
| 2 | AdaBoostM1 with IBK | 0.94 | 0.01 | 0.93 | 0.94 | 0.93 | 0.93 | 0.97 | 0.911 | 0.98 |
| 3 | AdaBoostM1 with Decision Table | 0.95 | 0.01 | 0.92 | 0.95 | 0.92 | 0.9 | 0.92 | 0.78 | 3.03 |
| 4 | AdaBoostM1 with Decision Stump | 0.47 | 0.18 | 0.92 | 0.45 | 0.92 | 0.9 | 0.86 | 0.33 | 0.04 |
| 5 | Attribute Selected Classifier with Bayes Net | 0.95 | 0.01 | 0.92 | 0.95 | 0.92 | 0.9 | 0.97 | 0.9 | 0.63 |
| 6 | Attribute Selected Classifier with IBK | 0.94 | 0.01 | 0.92 | 0.94 | 0.92 | 0.9 | 0.97 | 0.91 | 0.48 |
| 7 | Attribute Selected Classifier with Decision Table | 0.95 | 0.01 | 0.92 | 0.95 | 0.92 | 0.9 | 0.99 | 0.9 | 0.54 |
| 8 | Attribute Selected Classifier with Decision Stump | 0.45 | 0.18 | 0.92 | 0.45 | 0.92 | 0.9 | 0.86 | 0.33 | 0.47 |
| 9 | Bagging with Bayes Net | 0.95 | 0.01 | 0.92 | 0.95 | 0.92 | 0.9 | 0.98 | 0.91 | 0.33 |
| 10 | Bagging with IBK | 0.94 | 0.01 | 0.94 | 0.94 | 0.94 | 0.93 | 0.9 | 0.91 | 0.08 |
| 11 | Bagging with Decision Table | 0.95 | 0.01 | 0.92 | 0.95 | 0.92 | 0.9 | 0.99 | 0.9 | 3.59 |
| 12 | Bagging with Decision Stump | 0.95 | 0.01 | 0.92 | 0.95 | 0.92 | 0.9 | 0.99 | 0.9 | 3.59 |

The above table clearly represents the various metrics of this research work TP Rate, FP Rate, Precision, Recall, F Measure, MCC, ROC, PRC, and Time taken for build each model.

## IV. CONCLUSION

Main focus of this research work was finding best model over the data from Academic social network dataset i.e., Domain Paper Author dataset. The best classifier is AttributeSelectedClassifier with DecisionTable when compare with other models. The low accuracy have observed under AdaBoostM1 with DecisionStump and AttributeSelectedClassifier with DecisionStump. The optimal classifier that was observed when compared with the performance of other classifiers that were selected for this experiment.

# REFERENCES

[1] Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain Collaboration Recommendation. In Proceedings of the Eighteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2012).

[2] http://hanj.cs.illinois.edu/pdf/icdm12_hkim.pdf

[3] J. M. Hofman, A. Sharma, and D. J. Watts, "in Social Systems," vol. 488, no. February, pp. 486–488, 2017.

[4] http://arnetminer.org/lab-datasets/crossdomain/

[5] https://ieeexplore.ieee.org/document/7824855

[6] Y. Chen et al., "Delineating urban functional areas with building-level social media data: A dynamic time warping (DTW) distance based k-medoids method," Landsc. Urban Plan., vol. 160, pp. 48–60, 2017.

[7] S. Peng, G. Wang, and D. Xie, "Social Influence Analysis in Social Networking Big Data : Opportunities and Challenges," IEEE Netw., pp. 12–18, 2016.

[8] D. Gubiani and M. Pavan, "From trajectory modeling to social habits and behaviors analysis," in Studies in Systems, Decision and Control, vol. 66, 2017, pp. 371–385.

[9] http://www.cs.waikato.ac.nz/ml/weka/

[10] https://meshb.nlm.nih.gov/MeSHonDemand

[11] N. Nai-Arun and R. Moungmai, "Comparison of Classifiers for the Risk of Diabetes Prediction," Procedia Comput. Sci., vol. 69, pp. 132–142, 2015.

[12] https://edlab.tc.columbia.edu/blog/13139-Domain-Modeling-with-LDA-in-NLP-data-mining-in-Pressible

[13] A. Kaur and A. Datta, "A novel algorithm for fast and scalable subspace clustering of high-dimensional data," J. Big Data, vol. 2, no. 1, 2015.

[14] M. A. jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," Procedia Technol., vol. 10, pp. 85–94, 2013.

[15] M. T. Khan, M. Durrani, S. Khalid, and F. Aziz, "Online Knowledge-Based Model for Big Data Domain Extraction," Comput. Intell. Neurosci., vol. 2016, 2016.

[16] https://www.analyticsvidhya.com/blog/2018/10/stepwise-guide-topic-modeling-latent-semantic-analysis/