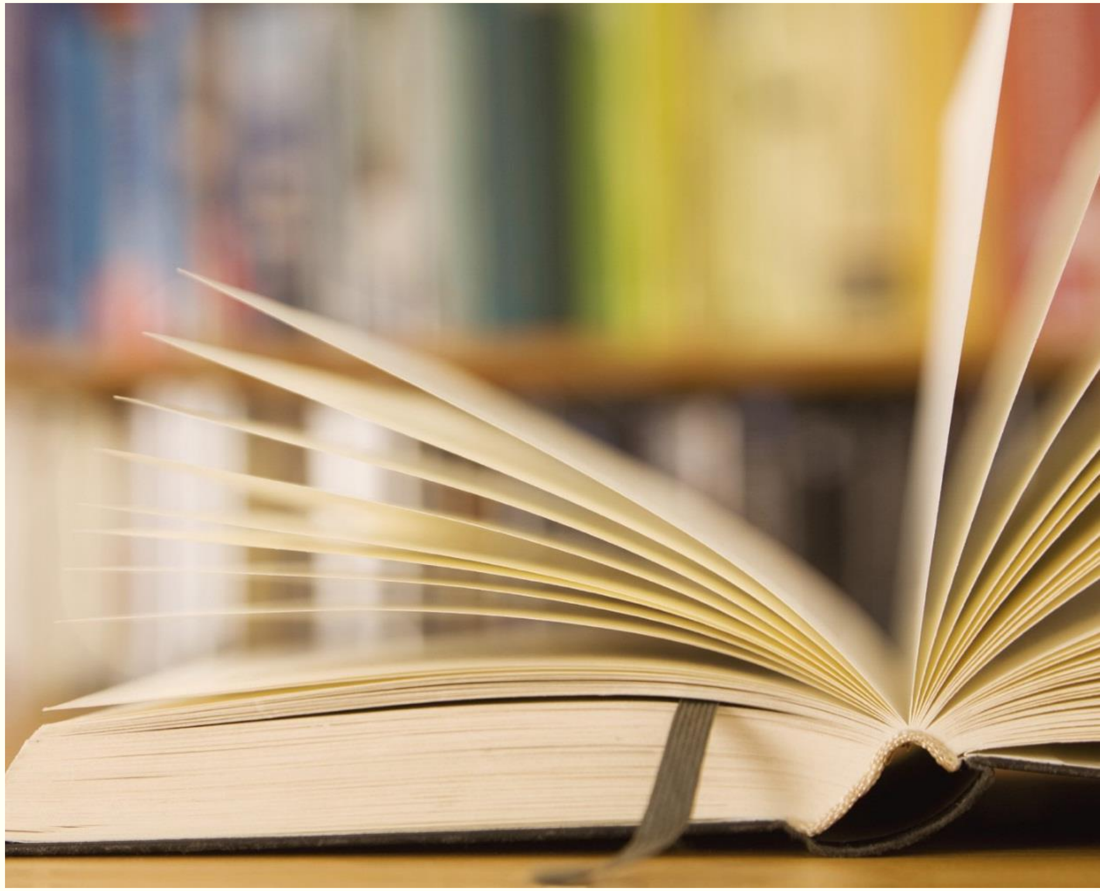
TOPIC CLASSIFICATION

Hybrid feature selection model using PSO and MLP



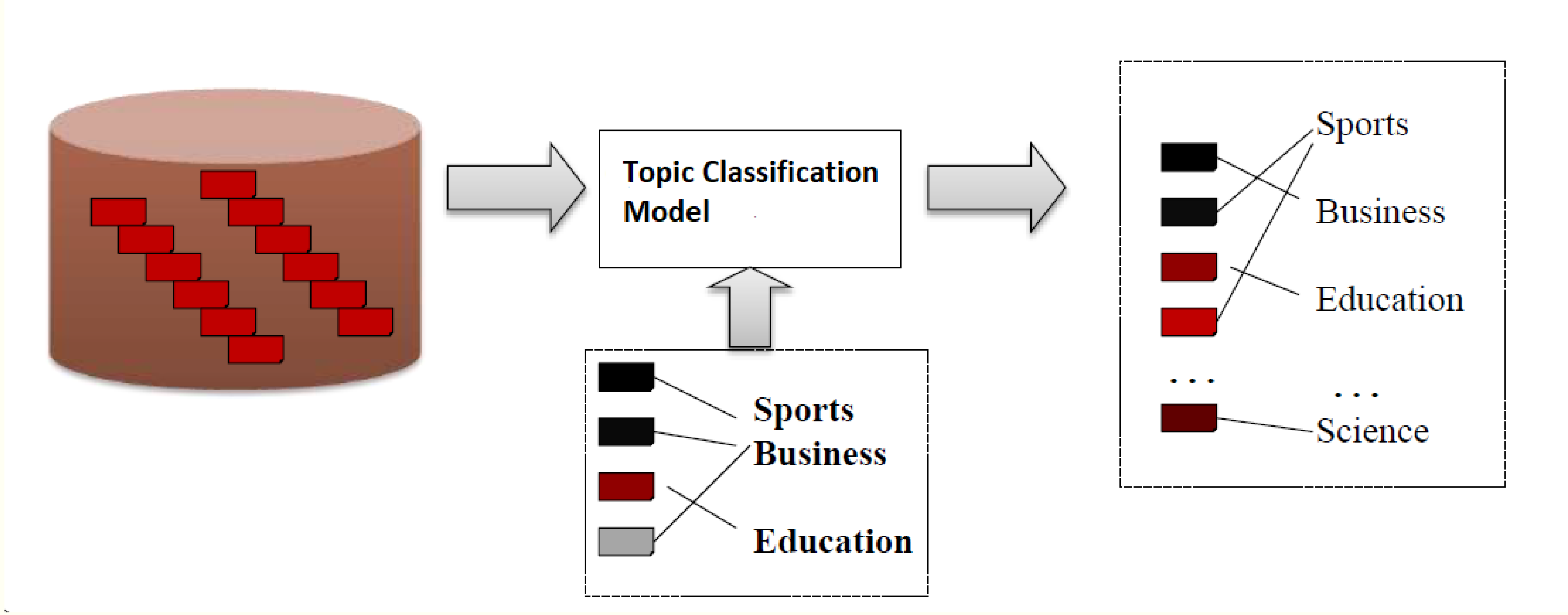
Karthikranjan Kunchum Satheesh

Motivation

- Organizing News feed data based on topics (Crime investigation's or law maker's)
- Analyzing problems in log-files (space center's)
- Categorize domain specific academic Journals/Articles
- Finding topic of interest Social media (Cambridge Analytica scandal)

Introduction

- ❑ To categorize the documents into pre-defined classes
- ❑ Helps to uncover hidden patterns from a huge pile of new documents
- ❑ A standard classification (supervised learning) problem



Research Problems

- Representing text is usually High-dimensional

  “Curse of Dimensionality”

- Optimal feature subset selection

  “Selecting the most relevant ones”

- overfitting of data

  “Fine tuning classifier”

Research Question

“This research investigates whether the proposed hybrid feature selection method can reduce the size of features and improve the classification accuracy of topic classification model with respect to news articles”

Related work

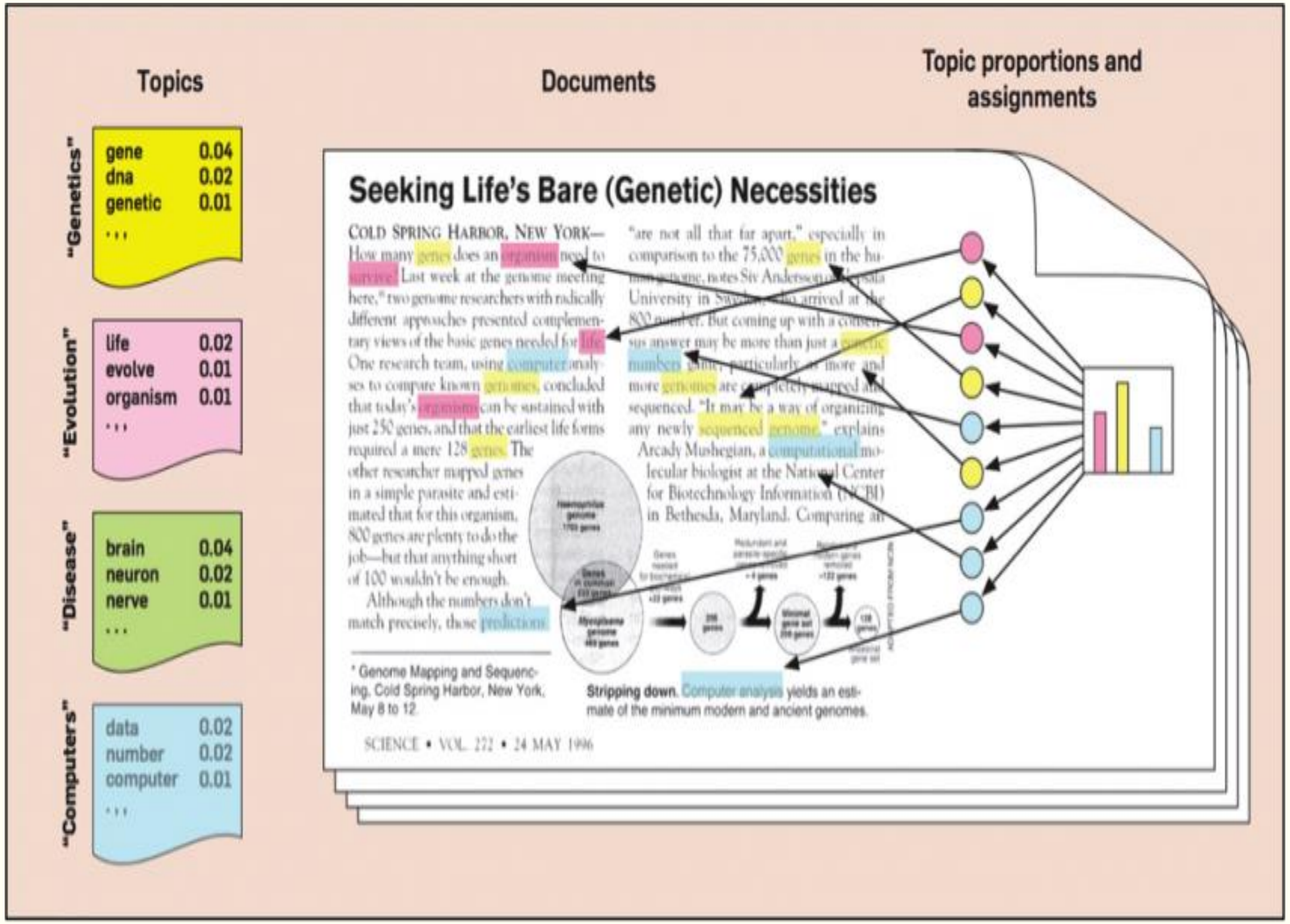
Search based techniques: "Bag-of-words" model approach

Feature Selection methods

- **Filter methods:** IG, Chi-Square, Tf-idf
- **Wrapper methods**: GA, PSO, AFSA.
- **Hybrid methods:** Trade-off between filter and wrapper methods

Classifiers: Naïve Bayes, SVM, MLP

Applications:
- Text classification (Labani et al. (2018))
- Gene Classification (Chuang et al. (2016))
- Image features extraction (Zhang et al. (2017a)
- Spam message filtering (Hu et al. (2015))

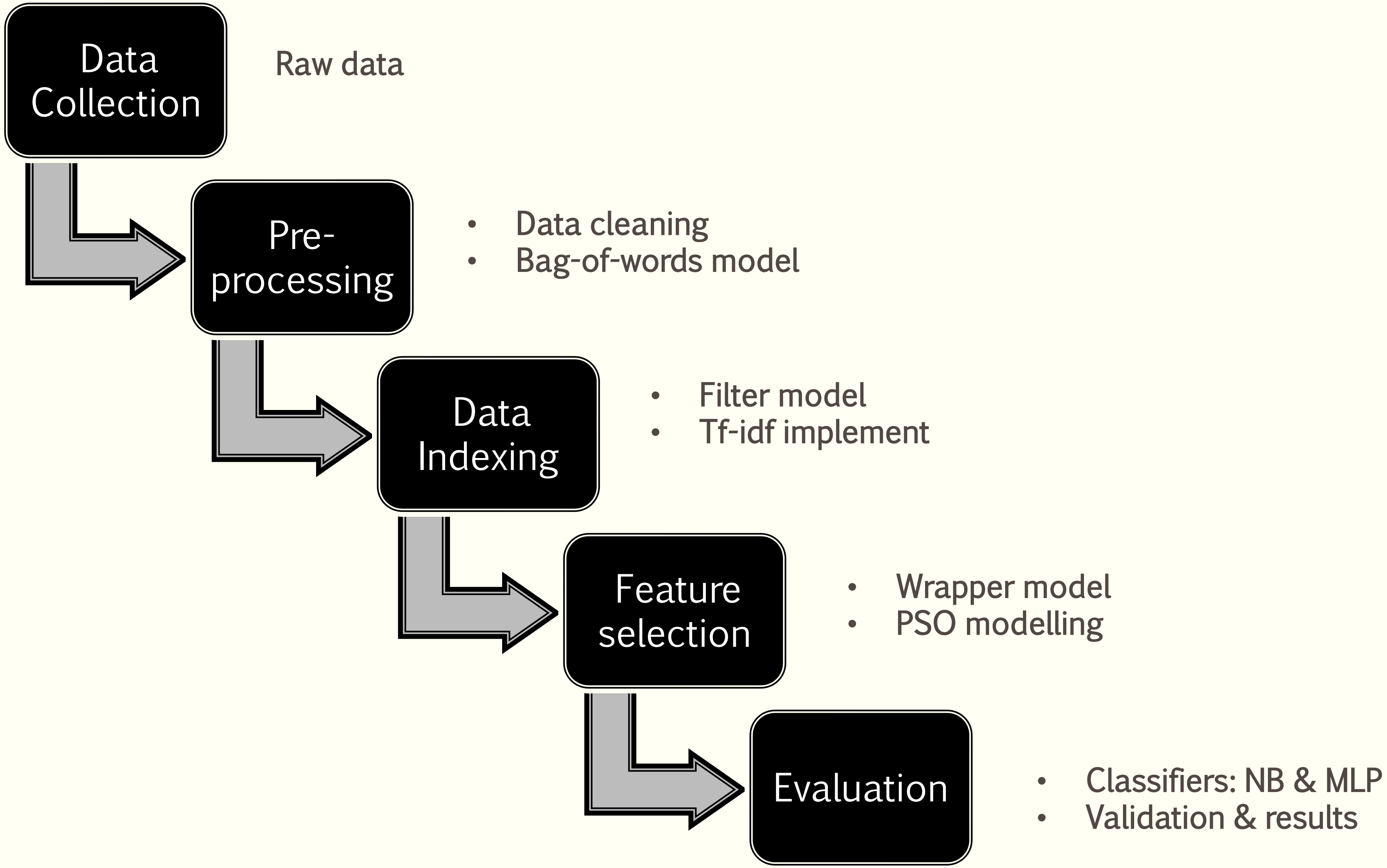


Topics
Documents
Topic proportions and assignments
"Genetics"
gene 0.04
dna 0.02
genetic 0.01
...
"Evolution"
life 0.02
evolve 0.01
organism 0.01
...
"Disease"
brain 0.04
neuron 0.02
nerve 0.01
...
"Computers"
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Objectives

- Information retrieval from large collection of unstructured data
- Identify the feature subset that are most useful for separating different categories
- To improve the classification accuracy on training data by optimally combining the features subset.
- The trained classifier can be applied to a test document to predict the most likely category.

Methodology



Data Collection

Raw data

Pre-processing

- Data cleaning
- Bag-of-words model

Data Indexing

- Filter model
- Tf-idf implement

Feature selection

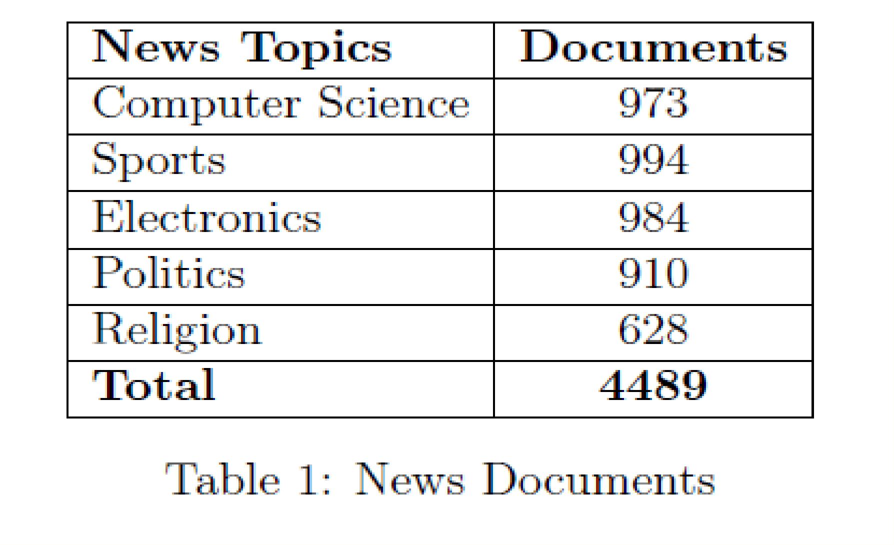
- Wrapper model
- PSO modelling

Evaluation

- Classifiers: NB & MLP
- Validation & results

Data Collection And Description

- Datasets from “**20NewsGroup**” text corpus repository was used for this research.
- Dataset Description:



News Topics	**Documents**
Computer Science	973
Sports	994
Electronics	984
Politics	910
Religion	628
Total	**4489**

Table 1: News Documents

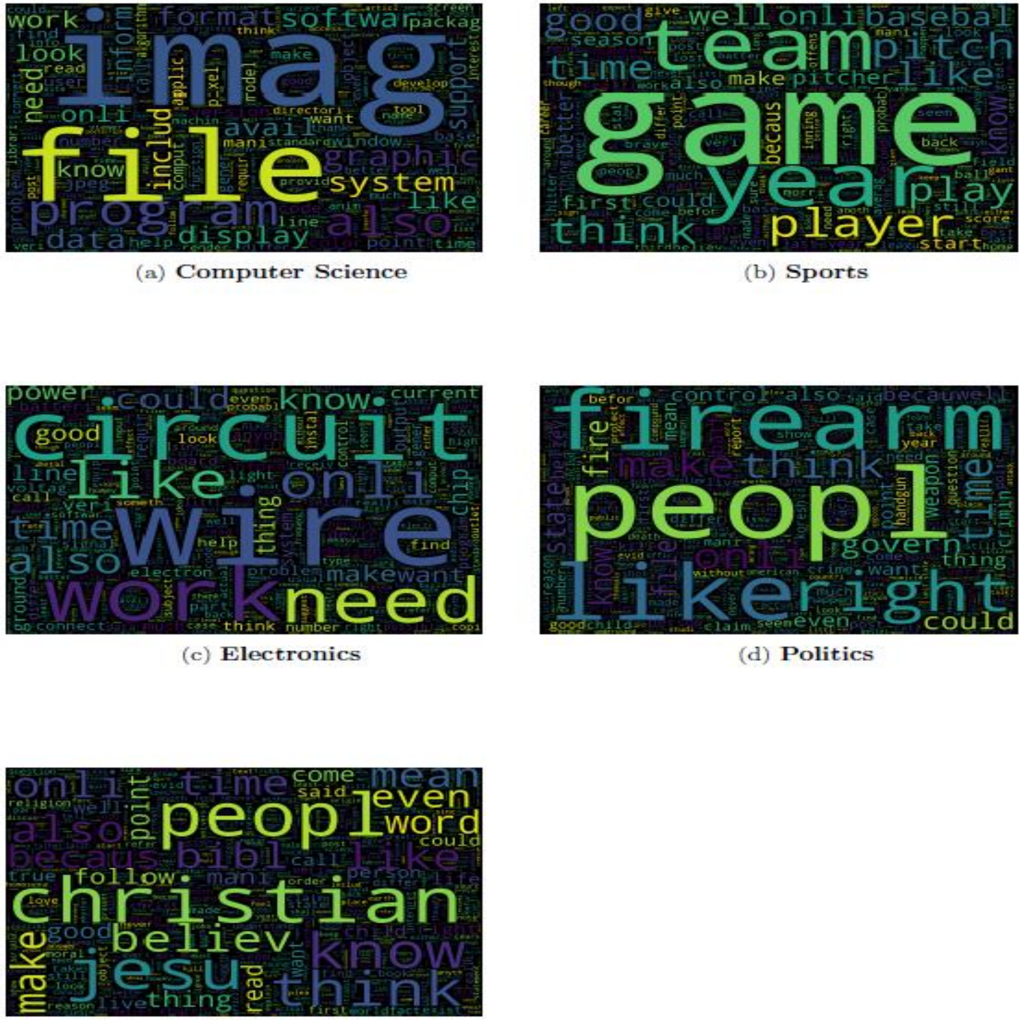
Data Pre-processing

The data cleaning includes

- Tokenization,
- Stop words removal,
- Stemming and Lemmatization.

Bag-of- words model creation

After pre-processing, the corpus data consist of 23406 input features/words.



(a) **Computer Science**

(b) **Sports**

(c) **Electronics**

(d) **Politics**

(e) **Religion**

Figure 2: **Wordcloud of News Topics**

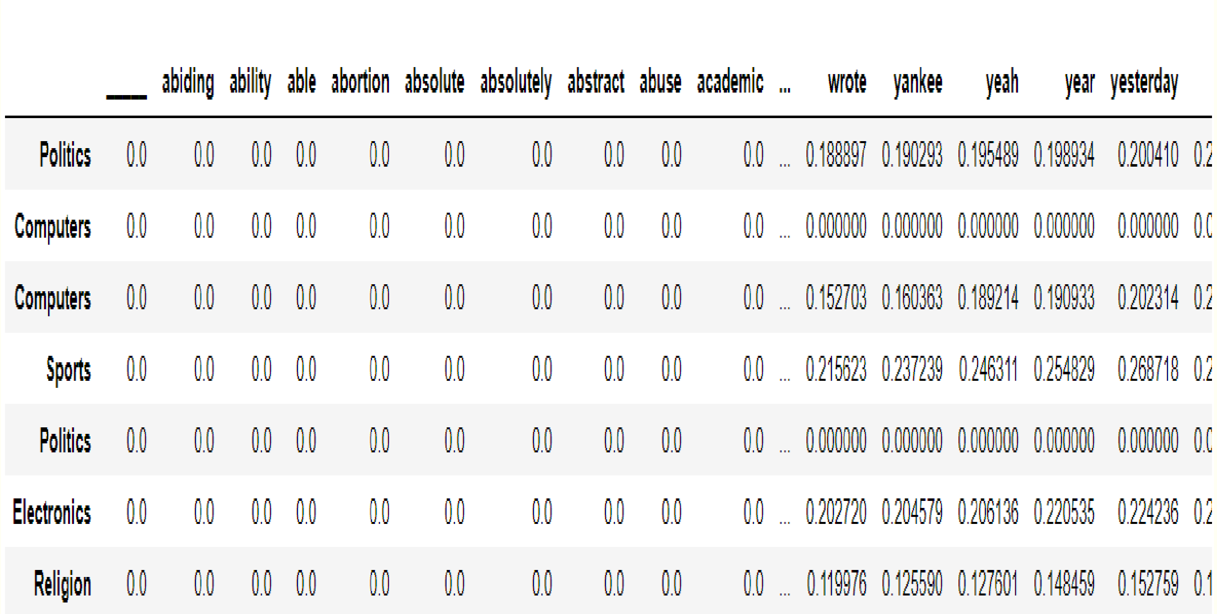
Data Indexing

Training and Testing

- No. of train Documents: 3591
- No. of test Documents: 898
- Total docs: 4489

Applying Tf-idf implementation (Filter method)

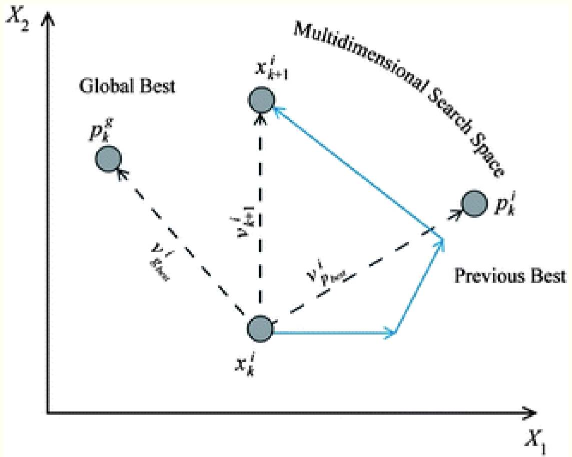
- Documents are represented as feature vectors
- A sequence of features and their weights

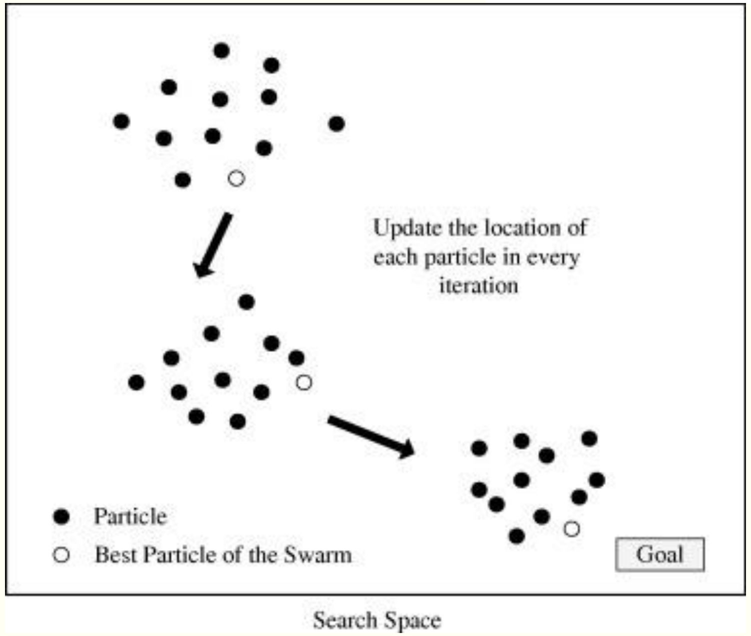


	____	abiding	ability	able	abortion	absolute	absolutely	abstract	abuse	academic	...	wrote	yankee	yeah	year	yesterday	
Politics	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.188897	0.190293	0.195489	0.198934	0.200410	0.2
Computers	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
Computers	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.152703	0.160363	0.189214	0.190933	0.202314	0.2
Sports	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.215623	0.237239	0.246311	0.254829	0.268718	0.2
Politics	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
Electronics	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.202720	0.204579	0.206136	0.220535	0.224236	0.2
Religion	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.119976	0.125590	0.127601	0.148459	0.152759	0.1

Particle Swarm Optimization

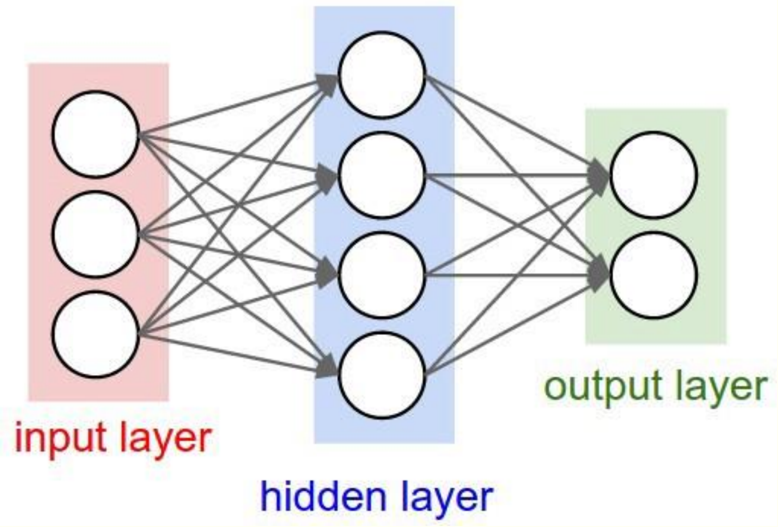
Particle swarm optimization (PSO) is a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. It solves a problem by having a population of candidate solutions, here dubbed particles, and moving these particles around in the search-space according to simple mathematical formulae over the particle's position and velocity (Kennedy, J., 2011).

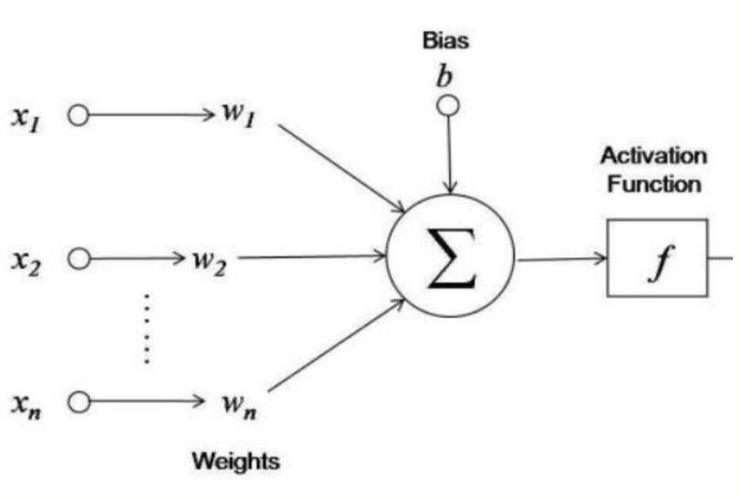



Update the location of each particle in every iteration
● Particle
○ Best Particle of the Swarm
Goal
Search Space

Multi-layer neural network

- The neural networks are trained by back propagation
- Modifying the weights in order to minimize the error
- Produce good results with complex datasets (high dimensional).
- Suitable for discrete data or text data
- Testing is very fast.

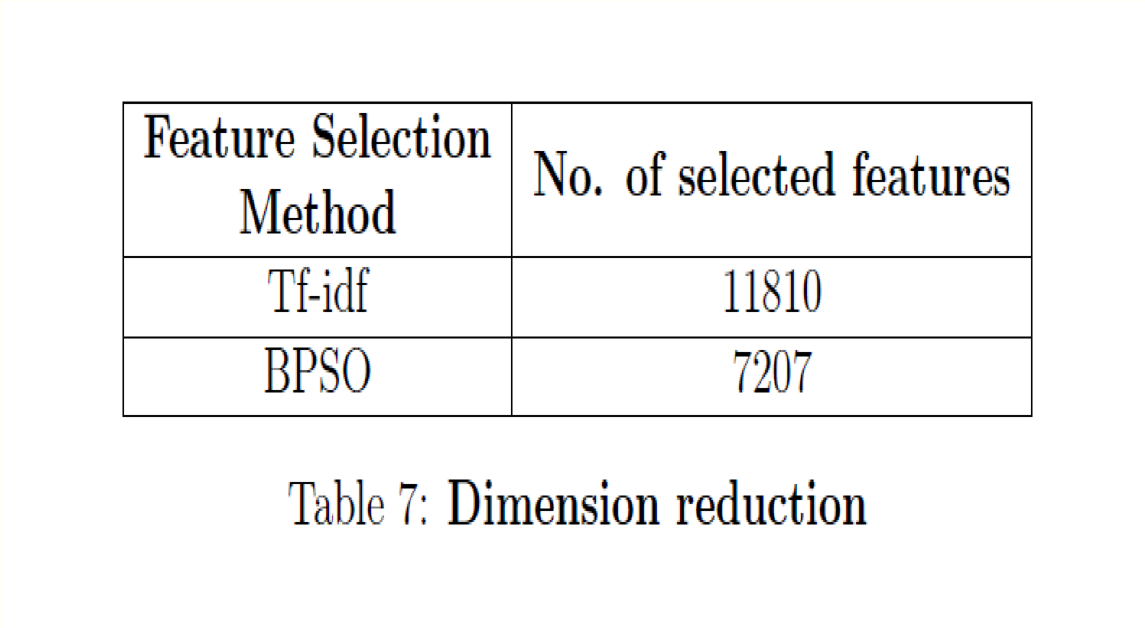




Evaluation criteria

1. 5-fold cross validation
2. **Precision:** High precision lesser the chance of false positive
3. **Recall:** Negative prediction rate
4. **F-measure:** Lesser the F-score higher the misclassification rate
5. **Accuracy:** To obtain optimum performance accuracy must be high

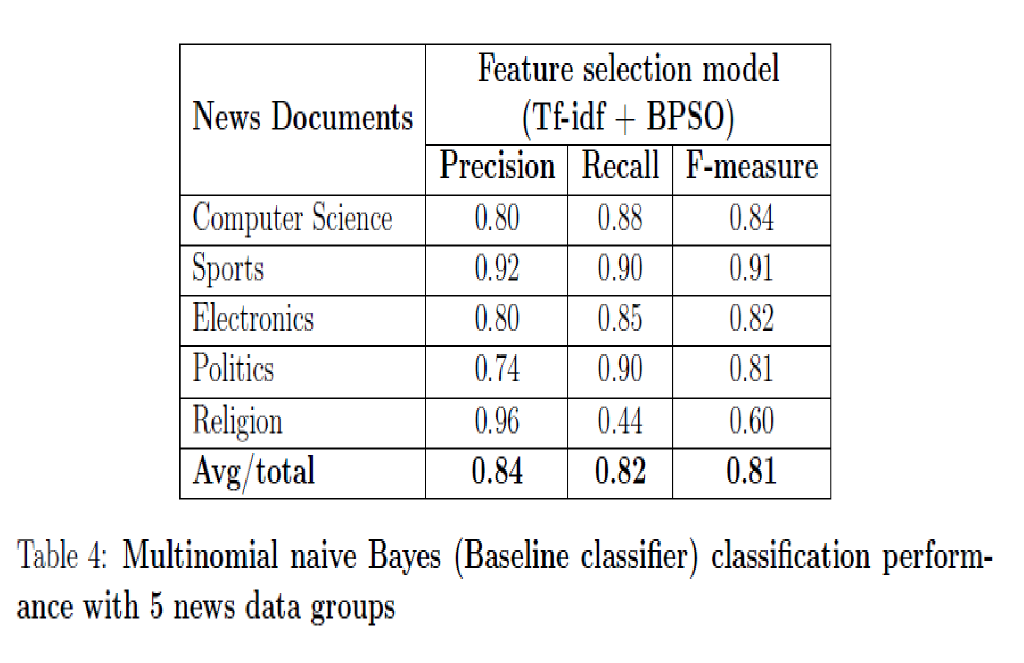
Dimensionality reduction



Feature Selection Method	No. of selected features
Tf-idf	11810
BPSO	7207

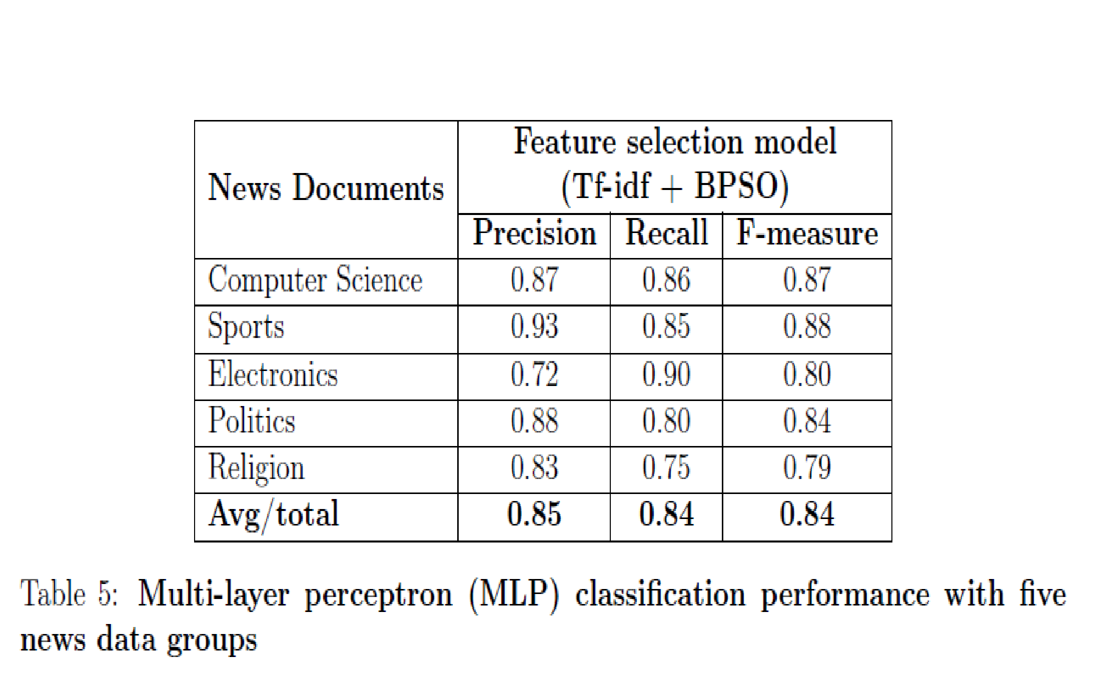
Table 7: Dimension reduction

Results Comparison



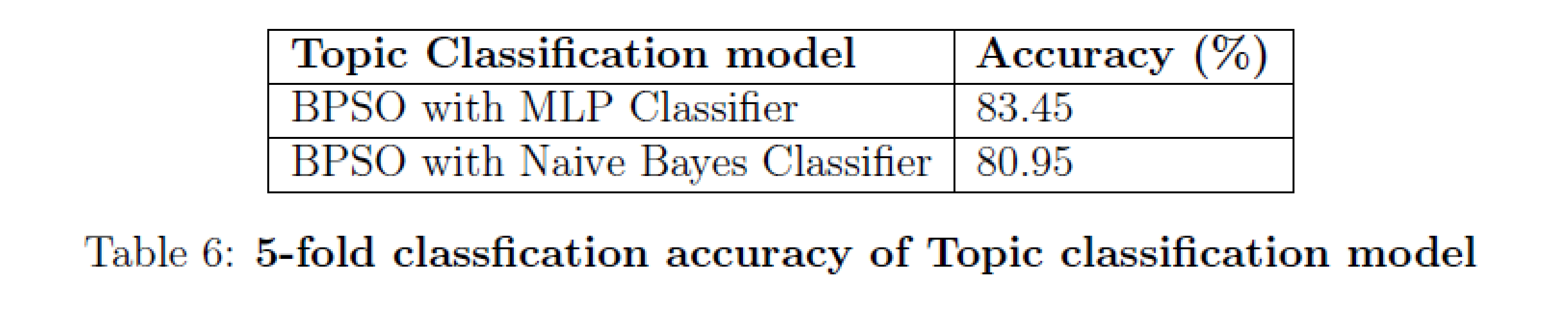
News Documents	Feature selection model (Tf-idf + BPSO)		
	Precision	Recall	F-measure
Computer Science	0.80	0.88	0.84
Sports	0.92	0.90	0.91
Electronics	0.80	0.85	0.82
Politics	0.74	0.90	0.81
Religion	0.96	0.44	0.60
Avg/total	**0.84**	**0.82**	**0.81**

Table 4: **Multinomial naive Bayes (Baseline classifier) classification performance with 5 news data groups**



News Documents	Feature selection model (Tf-idf + BPSO)		
	Precision	Recall	F-measure
Computer Science	0.87	0.86	0.87
Sports	0.93	0.85	0.88
Electronics	0.72	0.90	0.80
Politics	0.88	0.80	0.84
Religion	0.83	0.75	0.79
Avg/total	**0.85**	**0.84**	**0.84**

Table 5: **Multi-layer perceptron (MLP) classification performance with five news data groups**



Topic Classification model	**Accuracy (%)**
BPSO with MLP Classifier	83.45
BPSO with Naive Bayes Classifier	80.95

Table 6: **5-fold classfication accuracy of Topic classification model**

Conclusion

- Necessary to optimize the features and its subsets instead of just filtering based on frequency weights.
- Combination of feature selection methods Tf-idf and BPSO
- The results proved BPSO technique had better filtering process than Tf-Idf filter approach.
- Usage of neural network classifier like MLP helped achieve better classification when compared to multinomial naive Bayes

Future work

- Extend this work and test on large corpus of data and evaluate our model performance
- Using Hadoop distributed platform, to reduce processing time of BPSO and MLP and make classification task easier

References

- Kennedy, J., 2011. Particle swarm optimization. In Encyclopedia of machine learning (pp. 760-766). Springer, Boston, MA.

- Cao, J., Cui, H., Shi, H. and Jiao, L., 2016. Big data: A parallel particle swarm optimization-back-propagation neural network algorithm based on MapReduce. PloS one, 11(6), p.e0157551.

- Hu, H., Li, P. and Chen, Y. (2015). Biterm-based multilayer perceptron network for tagging short text, Cybernetics and Intelligent Systems (CIS) and IEEE Conference

- Zhang, Y., Gong, D.-w., Sun, X.-y. and Guo, Y.-n. (2017a). A pso-based multi-objective multi-label feature selection method in classification, Scientific reports 7(1): 376.

- Chuang, L.-Y., Ke, C.-H. and Yang, C.-H. (2016). A hybrid both filter and wrapper feature selection method for microarray classification, arXiv preprint arXiv:1612.08669

- Labani, M., Moradi, P., Ahmadizar, F. and Jalili, M. (2018). A novel multivariate filter method for feature selection in text classification problems, Engineering Applications of Artificial Intelligence 70: 25–37.

- https://hackernoon.com/challenges-in-deep-learning-57bbf6e73bb