Topic Classification: Hybrid Feature selection model using BPSO-MLP

MSc Research Project
Data Analytics

Karthikranjan Kunchum Satheesh
x17110254

School of Computing
National College of Ireland

Supervisor: Dr. Pramod Pathak
Dympna O'Sullivan
Dr. Paul Stynes

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Karthikranjan Kunchum Satheesh
Student ID:	x17110254
Programme:	Msc Data Analytics
Year:	2018
Module:	Research Project
Lecturer:	Dr. Paul Stynes
Submission Due Date:	13/08/2018
Project Title:	Topic Classification: Hybrid Feature selection model using BPSO-MLP
Word Count:	5092

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	15th September 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Topic Classification: Hybrid Feature selection model using BPSO-MLP

Karthikranjan Kunchum Satheesh
x17110254
MSc Research Project in Data Analytics

15th September 2018

Abstract

In this 21st century era of information and technology where our day to day life is filled with online activities which generate data by leaps and bounds. Majority of data being generated is of natural language/ text data format which becomes difficult to store. if this data is managed and analyzed properly has immense potential to make businesses more convenient. So, one must provide a way through to navigate and process this text data of reports for categorising the topics of interest. Since it would be a challenging task to handle high dimensional data and filter a relevant information, feature selection algorithms were used to enhance the efficiency of the topic classification and improve in a choice of selecting significant information. This study investigates the efficiency of the hybrid feature selection technique based on binary particle swarm optimization and evaluated with multi-layered perceptron to determine the quality of features. Experimental results show that the combination of multi-layered perceptron with B-PSO resulted in a good accuracy rate of 83% with a reduced number of features.

Keywords: MLP (multi-layer perceptron), B-PSO (Binary particle swarm optimization), feature selection.

1 Introduction

In today's fast-paced environment, the growth of internet information has enabled access to a vast amount of text data from internal and external sources. Much of this unstructured text is valuable in helping companies to discover the hidden patterns and in making strategic business decisions. Moreover, in reality, captured knowledge from the text-based information assets are used as actionable insights which are key for businesses. Especially the knowledge-driven industries such as life sciences, healthcare, financial institutions, government entities, space research centres etc. are in the process of maximize the value of information assets to find business insights. The huge amounts of information assets exist in the form of unstructured text corpuses such as scientific articles, healthcare records, space problem reports and news feed corpora. Among these, the most popular use case in organising and recommending similar news articles based on the specific categories (topic areas) on a news website. Currently, a news website which provides a variety of news articles a day to users on a broad range of topic areas. To increase the user engagement and time span on the site, maybe one of the solutions is to recommend similar articles in an organised way. So, one must provide a way through to navigate and process all these corpora of reports for identifying the most popular topics and trending themes to deliver a prescriptive analytical result that adds a value to the business. But this is all often poorly exploited. In order to achieve an effective classification model in processing and analysing the textual data by identifying relevant text features while reducing the access time. So, many researchers have promoted various topic classification techniques.

Topic classification is broadly used to categorize the text documents of predefined classes, this method helps to uncover a hidden thematic structure in a huge pile of documents. In the process of topic classification, documents are basically modeled as "bag of words" or vector space Sahon and McGill (1983), in which each word represented as features. The word vectors in the models are defined by values based on term frequency or term frequency-inverse document frequency which are interpreted as input features that classify the topics in analysis. Since it requires to process a huge amount of data and handle high dimensionality it would be a challenging task for classifying relevant documents in limited computational time. To overcome this problem employed feature selection methods for efficient text classification tasks.

Numerous studies on feature selection methods are focused to find the solution for high-dimensionality issues and maintaining the performance of the classifier. Filter methods mainly rely on the statistical score for ranking each feature by weight vector terminology. They are generally faster compared to wrapper methods but if the computed values are highly correlated causes feature redundancy which leads relatively poor performance. In contrary hybrid methods is a trade-off between filter and wrapper methods for achieving higher classification accuracy. The wrapper and hybrid methods are computationally expensive as it depends on the choice of learning classifier, especially used to reduce the dimensionality in large scale features space. Recently evolutionary algorithms like a genetic algorithm, particle swarm optimization, etc are quite popularly used as a feature selection technique which mainly relies on a classification algorithm.

Various studies demonstrate the efficacy of PSO in search of an optimal subset of features from the images Zhang et al. (2017a). As compared to other evolutionary algorithms, PSO is easier to implement and faster convergence to the solutions for obtaining relevant feature set. To handle high-dimensionality in features in less operation time choosing the finest classifier is the next most important step in text classification. So, using Neural network topology as combined with feature selection as shown significant performance with large-scale high dimension datasets Ye (2017). In this research, binary particle swarm optimization (BPSO) combined with multi-layer neural network (MLP) classifier is used to implement a feature selection method and used to evaluate the fitness of PSO for optimally selecting the subset of features. Here the accuracy and precision are key metrics for evaluating the performance of the classifier.

" 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."

The research process flow given below:

- Identify the subset of features that are useful for separating different categories.
- To improve the classification accuracy on training data by optimally combining the features subset.
- The trained topic classification model can be applied to a test document for predicting the most likely category.

The research paper further constitutes into following sections. Section 2 discusses regarding similar works on text classification and a brief review of feature selection techniques. Section 3 presented the proposed workflow process and techniques used in this research. Section 4 explains the process of implementation steps. Section 5 will be the result and discussion. Finally section 6 conclusion and future work.

2 Related Work

Numerous studies have been done for text classification analysis in a different domain using probabilistic/statistical models and search-based text classification models. Among these studies, search-based text classification models that are most popular since these models are automatically trained to identify the informative features while maintaining the performance of the classifier.

In most of the text classification tasks, documents are represented as a "bag of words" or vector space model Sahon and McGill (1983) where a set of documents are categorized into classes and words that represented as features. Further, the frequency of each word is used as input features for training a classifier. Although the text documents contain thousands of words/features which leads to high dimensionality problem for classification. So, the feature relevance or ranking indicates the importance of feature which is always necessary to predict the class labels and redundant features are defined based on the correlation among features. The aim of feature selection is to select the most relevant features subset with minimum redundancy.

Feature selection is the most important step carried out after pre-processing stage which involves the use of various categories of topics and words in the documents that are ranked as features. There are many feature selection methods which generate the value for each feature in a large vector space Sahon and McGill (1983). Traditional feature selection methods are classified into two types based on their selection strategy such has filter and wrapper methods. Filter methods mainly rely on the statistical score for ranking each feature by weight vector terminology. They are generally faster compared to wrapper methods but if the computed values are highly correlated causes feature redundancy which leads relatively poor performance. Some of the ranking metrics such as Information gain, Chi-squared ratio, Document frequency (Df), term frequency (Tf), and Tf-IDF (term frequency and inverse document frequency).

2.1 Filter based features selection methods

Labani et al. (2018) addressed some of the filter methods like information gain, Gini index MRDC (Multivariate Relative Discrimination Criterion) for features selection. In which obtained result were computationally faster but failed to handle redundant features. The classification performance of information gain and MRDC on news articles performed well for a greater number of features with 35% and 34% precision respectively. As researchers suggest Forman (2003) that usage of Tf-Idf takes care of parameters for selecting the predominant features in the analysis but still lacks in handling the dimensionality problem. The resulted features are quite large which impacts the performance of the classifier. Recent studies also suggest that usage of the wrapper method produces the best feature subset with underlying classifier algorithm. They usually have superior classification accuracy than filter methods but they need more computational power and rely on a performance of classifier for selecting the subset of features.

2.2 Wrapper-based features selection methods

Feature selection is a challenging task since the available features in a search space grow exponentially with an increase in the size of the dataset. Due to global search ability, Evolutionary algorithms are widely used as feature selection methods for solving complex problems easily in polynomial time. Evolutionary techniques use wrapper-based feature selection strategy where it mainly relies on the underlying classifier such has a Genetic algorithm, artificial bee colony optimization, Particle swarm optimization, ant colony optimization and artificial swarm optimization.

The genetic Algorithm most popular evolutionary technique, Labani et al. (2018) applied on scientific articles in segregation of topic from Persian text documents. The genetic algorithms produced better results for all documents size as compared to Tf-idf method. Both of the methods reported the similar precision of 50% for larger document size. Pudaruth et al. (2017) proposed a random walk algorithm with SVM classifier for categorising the legal documents. A comparison was made between the genetic algorithm and simulated annealing feature selection techniques for identifying 6 different topics which shown better results for a less sampled data as compared to the proposed method. In another hand Nalluri et al. (2017) proposed a wrapper approach of AFSO feature selection on medical datasets for classifying the disease diagnosis. This method was tested on imbalanced data with highly correlated features of multiple classes. Using SVM classifier they achieved a better classification accuracy of ranging 3.15% to 22.8% with limited features subset.

PSO is introduced by Ratnaweera et al. (2004) which is search based optimization technique that initializes a population of particles in an N-dimensional vector space. In addition Zhang et al. (2017b) addressed binary PSO for solving multi-label classification problems such as emoticons, images feature from the trained dataset. However binary PSO variant is a most preferred technique for discrete values ranging from 0 to 1. Currently, most of the studies focus on modifying BPSO for obtaining a specific solution for a problem. The PSO algorithm was improvised with a mutation operator for efficient local learning strategy in finding optimal solutions in a single iteration. The desired results are evaluated using SVM and KNN (K-nearest neighbor) classifiers.

2.3 Hybrid feature selection methods

Chuang et al. (2016) presented a study on gene classification problem by integrating filter method (information gain) and wrapper method (PSO) for high classification accuracy. Exploratory results show that employing proposed method with SVM classifier which yielded better accuracy rate of 87.1% with fewer gene features. In another hand, the Abualigah et al. (2017) proposed particle swarm optimization and genetic algorithm with new weighing terminology (LFW) for improving the selection of features in clustering documents. Genetic algorithm bound to have overfitting with trained data due to long running times. In contrast, modified PSO yielded significant results for document clustering. The limitation of the genetic algorithm is due to crossover and mutation factors which lead to devolvement of PSO for determining the global optimum solution.

Some of the recent studies suggest that choosing an efficient classifier is the most important step for text classification. Most of the traditional classifier such as Naive Bayes, decision trees and support vector is inefficient in handling a large volume of data which requires extensive memory usage. These complex tasks can be addressed using neural networks that are applied in various domains such as voice recognition, image processing, and text processing. The most relevant work conducted by Souza et al. (2018) used binary PSO for selecting the most significant features from the voice detection system. The fitness function is calculated using a multi-layered neural network based on the accuracy rate. In addition to that Brito et al. (2017) , the author recommends usage of feature selection technique more prominent with the neural network since it's hard to train the large set of features and may prone to overfitting. Furthermore Hu et al. (2015) proposed a short text tagging model based on the word co-occurrences (Bi-term) from the document. The collected data was linearly separable and the text was tagged with different categories using Multilayer neural network which achieved good accuracy rate.

In previous studies described, they tested either filter or wrapper-based feature selection techniques for text classification problem. But in this research, a hybrid feature selection of filter and wrapper approach is implemented for classifying the topics in news documents. Hybrid approaches yielded significant results for gene classification problems, for example, but have not been widely applied to text. It motivates us to develop a hybrid feature selection model for better text classification results.

3 Methodology

This research focuses on the classification of the relevant topics from an unstructured text data in the process of aiding businesses decision support system. The proposed research method is on the usage of binary particle swarm optimization as feature selection and evaluated using two classifiers such has multinomial Naive Bayes and Multi-layer perceptron for classification of documents. Fig 1 shows the workflow of the proposed method for topic classification. A KDD approach was followed in the entire process.

The collected data from the 20newsgroup corpora having various categories with a huge set of classes in documents Lang (2015) . This corpus contains a collection of 18k news feed articles of 20 different categories/topics that appeared in Usenet newsgroup collection. Two of the topics are distinct and the rest are closely related to each other.

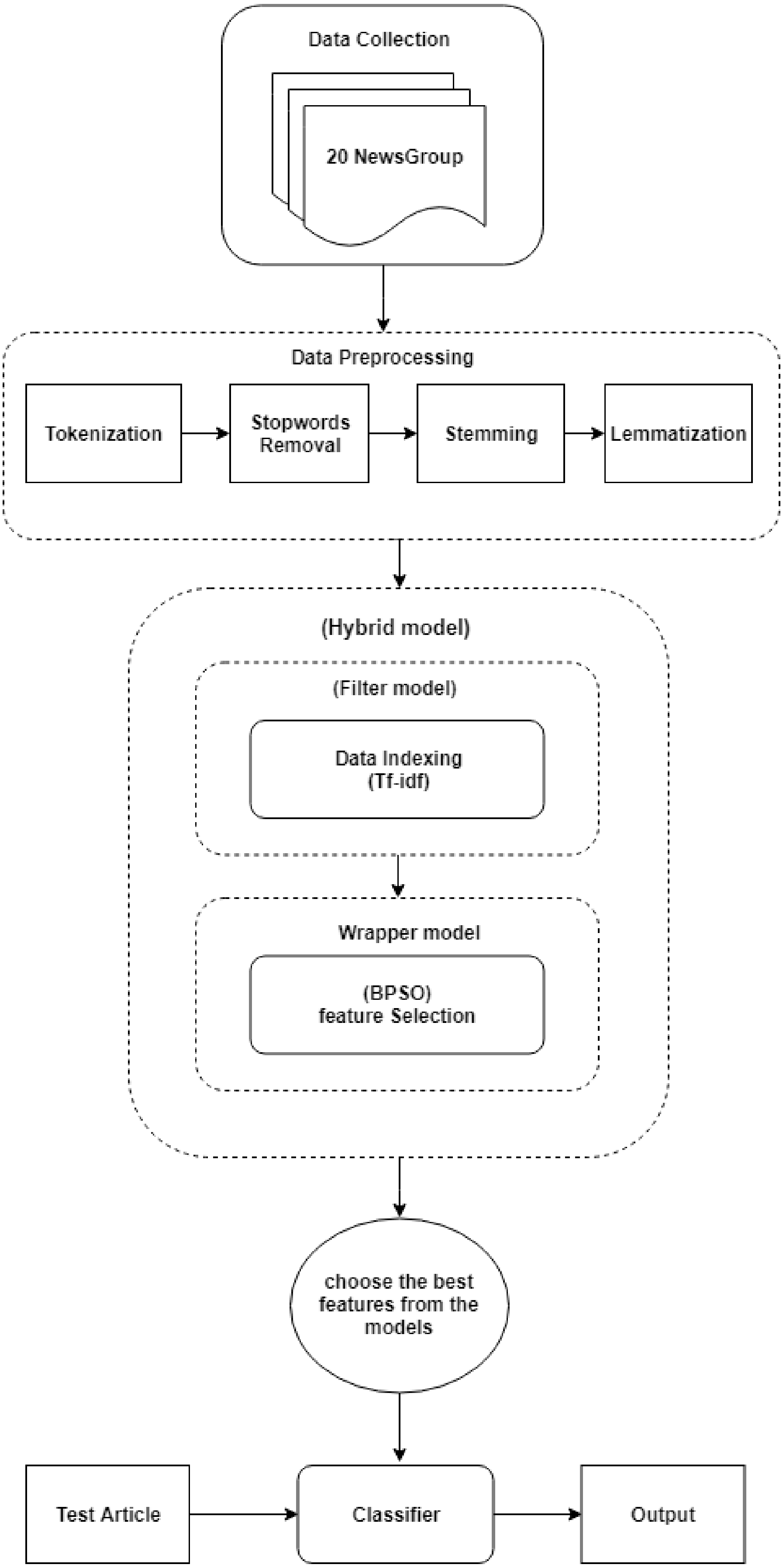


Figure 1: **Proposed workflow**

Apart from this, corpora have a huge set of lexicons and less usage of rhetorical writings.

The collected corpus data is of textual format which is inconsistent to apply machine learning models for text classification, therefore the need to undergo pre-processing i.e. cleaning, removal of noise, unwanted and junk data. Transformation is done in terms of continuous text to meaningful words(features). The following process is of

1. **I**dentifying and categorizing the features as "Rankings" based on the occurrence of the particular feature in the corpus. The process of ranking the features can be achieved using term frequency and inverse document frequency (Tf-idf) filter-method approach

2. **A**ll these input features of the data will not be important, as this will reduce the overall performance of the proposed method. Feature selection is used to find out optimal set of "feature" which can be further useful in segregation of topics. Using binary particle swarm optimisation obtained an optimal subset of features with low dimensionality in feature space.

3. **C**ompared and evaluated the selected set of features from the filter method (Tf-idf) and proposed hybrid feature selection method (Tf-idf and BPSO) using two classification algorithms i.e. multinomial naïve Bayes and multi-layer perceptron. Multinomial naive Bayes used as baseline classifier.

Finally, considering accuracy, precision, recall and f-measure as evaluation metrics, the calibrations and parameters are finely tuned and adjusted to improve classification performance.

4 Implementation

In this section, the experiments are performed to enhance the classification performance using the proposed feature selection method (BPSO) with two different classifiers such as Multinomial naive Bayes and multilayer perceptron. These experiments were conducted on the 20-Newsgroups dataset. All these experiments were set up in windows 7 installed machine, having a core i7 processor, 8GB RAM and NVIDIA GPU unit. The technological resources used in this research are python 3.6 IDE from Anaconda distribution, Jupyter notebook, and Python programming language. Additional software installation is required Tensor flow and Keras for running neural nets in CPU/GPU mode. In the following sections process of implementation is explained in detailed.

4.1 Data Retrieval

The major hurdle in supervised learning is collecting the multi-labeled text documents with predefined classes. since this research is about the topic classification of documents, so collected data from the 20newsgroup corpora having 5 categories with a huge set of classes in documents Lang (2015) . Data were automatically retrieved from the 20Newsgroup website which contains a corpus of 4489 text documents. For retrieving the data, used Scikit-learn package which as pre-installed fetch20newsgroup API class that can fetch the data from the corpus repository. Finally, the collected documents as 4489 rows of 5 set of classes were stored as shown in below table 1.

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

Table 1: News Documents

4.2 Data Pre-Processing

Data Pre-processing plays a crucial role and the most time-consuming stage of the entire process. As the collected raw data from the corpus will be inconsistent, redundant and noisy it needs to be cleaned and processed for further analysis. The data cleaning includes tokenization, stop words removal, Stemming and Lemmatization. The tokenization processes the document into small chunks of words which are referred to as tokens and removes special characters in the document. Redundant words in the document are removed by using standard library of stop words in NLTK package *NLTK 3.2.5 Documentation* (2015) . Lower/upper cases, suffixes, prefixes are normalized through the stemming process. Lemmatization groups and considers a canonical form of stem words as one single entity. After pre-processing, the corpus data consist of **4489** documents and **23406** input features. To interpret the corpus data with different topic related features, created wordcloud which gives a visual representation of text features with corresponding topics as shown in the figure 2.

4.3 Filter based feature selection

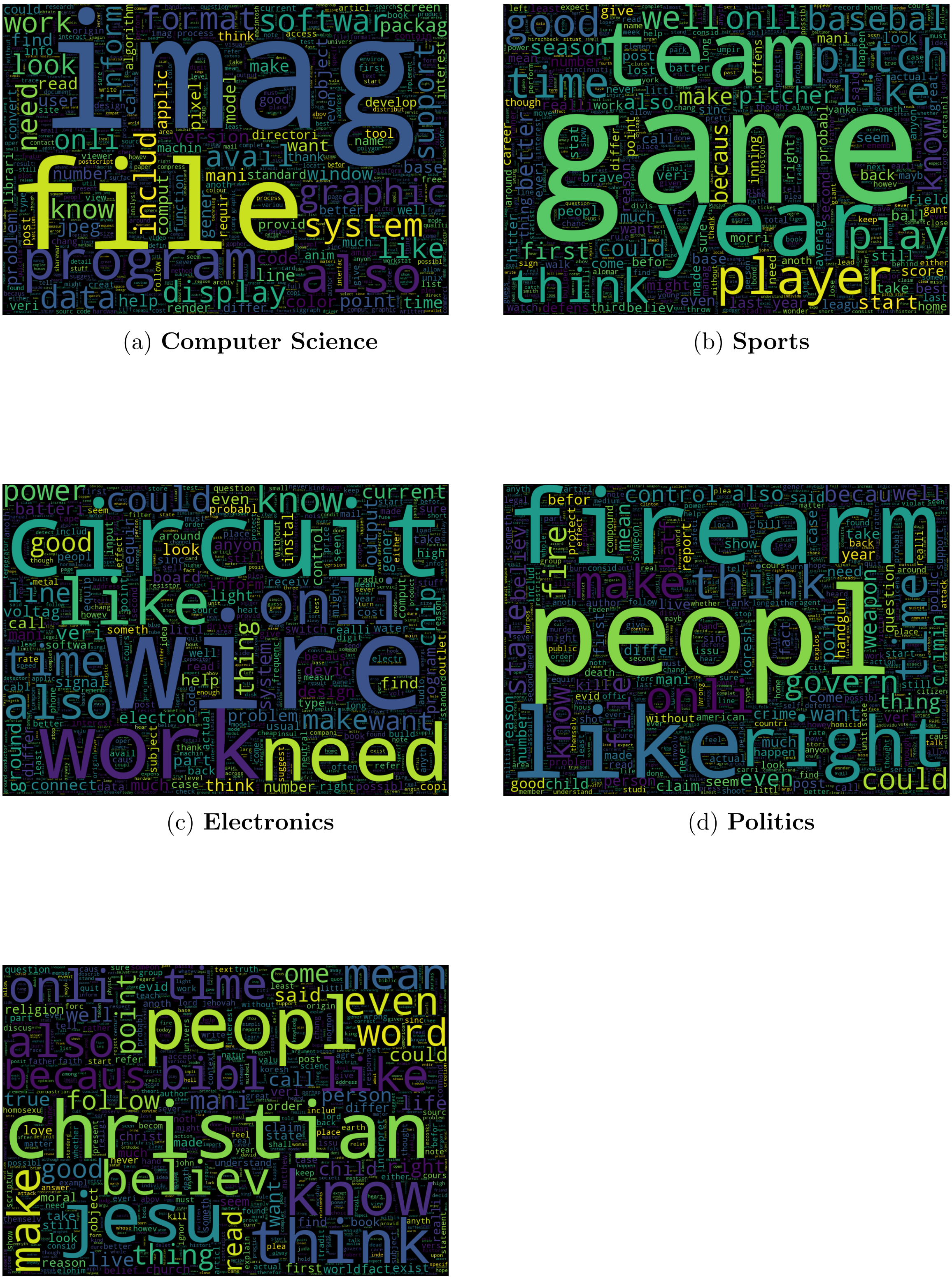
The last step of pre-processing is a translational stage at which the human-readable language is translated into machine-readable language, the i.e. text is transformed into numerical vectors. The indexing step is done with the help of a filter method Tf-IDF weighting i.e. (Term frequency-inverse document frequency) Forman (2003) . This method is executed in a sequence of steps

Step 1: Initially, with the help of tokenization result, created a “bag of words” model by calculating the frequency of each token using “count vectorizer” class.

Step 2: Since each token cannot be deemed as a prominent or useful one, we undertake the process of finding out distinct tokens representing a useful topic(s).

Step 3: This step includes measuring IDF - the frequency of word occurring in all documents.

Step 4: Finally, we calculate the weight of each token as a product of Tf and IDF using this we can calculate the value of each word/feature.



(a) **Computer Science**

(b) **Sports**

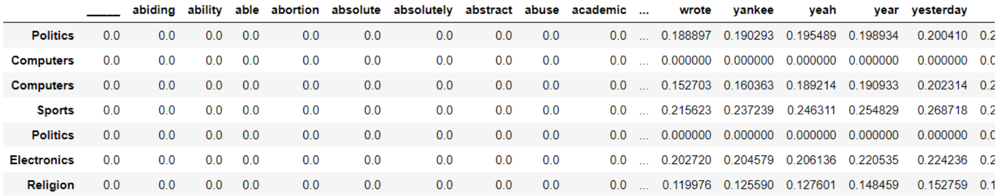
(c) **Electronics**

(d) **Politics**

(e) **Religion**

Figure 2: **Wordcloud of News Topics**

This entire process is carried out using Tf-idf transformer a pre-installed algorithm in the NLTK package *NLTK 3.2.5 Documentation* (2015) . At the end of this Tf-idf process, the weighted values are assigned to each token and uploaded into the data frame which comprises features and topic data. So, the final dataset contains 4489 rows as topic data and 11810 columns represent features as shown in the fig 3.



	____	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.C
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.C
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

Figure 3: **Dataframe**

4.4 Wrapper-based feature selection: Binary particle swarm optimization

Since a high number of input features in the pre-processed data is more complex to train the model, in order to reduce the dimensionality and improve classification performance proposed BPSO feature selection technique and evaluated using multi-nomial naive Bayes and multi-layer perceptron classifiers. Ratnaweera et al. (2004) Initially, the BPSO creates a population of particles randomly in feature space. Each particle covers up with possible solutions at each iteration by updating their position and velocity. After every iteration two best solutions (fitness) are calculated called "pbest" and "gbest". The movement of each particle is tracked by the coordinates in the problem space, which are associated with the best solution that found in the past. The idea is that more and more particles will eventually move towards areas where better solutions are found and that the population will eventually converge to the optimal value. The performance of BPSO can be improved by setting optimized PSO parameters as shown in the below table 2. Here the fitness of BPSO is evaluated using multi-layered perceptron and multinomial naive Bayes classifier. The feature selection models were implemented using "Pyswarm" toolkit library *Pyswarm* (2015) .

No. of Particles	35
Stop Criteria (Iterations)	10
Acceleration factors (C1 = C2)	2
Fixed inertia factor (w)	0.9
Input features to BPSO	11,810

Table 2: **BPSO Parameters**

4.5 Classification Model:

In this stage, a resulted subset of features selected from BPSO is used to evaluate the classification performance using multinomial naive Bayes and multi-layer perceptron classifiers. Multinomial naive Bayes is used as baseline classifier to compare the performance of classification results. In case of multi-layer perceptron, the size of the feature subset which relates the number of inputs to the input layer. The number of neurons on the hidden layer is computed by taking average count of neurons on input and output layers. Akilandeswari and Nasira (2015) All of the weights initialized with random values and updated using an adaptive moment estimation algorithm (Adam) back-propagation method. There were 10 epochs employed in this research. Table 3. shows the parameters used for the multi-layer neural network.

No. of layers	3
No. of hidden layers	1
No. of neurons in the hidden layer	1024
No. of neurons in the output layer	512
No. of epochs	10
Activation function	sigmoid
Optimisation function	adam

Table 3: **MLP Parameters**

5 Results and Discussion

In this section, the test results of feature selection models using the proposed hybrid method (Tf-idf + BPSO) were evaluated based on classification performance of 5 multi-category news articles. After the feature selection mechanism, the selected feature subsets are evaluated using two generic classification algorithms i.e. multinomial naive Bayes and multi-layer perceptron (MLP). Here the multinomial naive Bayes used as baseline classifier. The evaluation metric used in this experiment are accuracy, precision, recall, and f-measure.

In Table 1 and Table 2 shows the comparison of classification performance of Multinomial naïve Bayes and multi-layer perceptron classifier. The precision and recall values for both the classifiers resembled similar results for all news data groups. Since the sampled data as fewer categories for classification and to get accurately predicted values, K-fold cross-validation technique is applied to cross validate the classification results.

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**

In table 6 shows the 5-fold classification accuracy of proposed feature selection model evaluated using naive Bayes and MLP classifiers. As observed in table 6 , comparatively there is a marginal difference in accuracy between the classifiers. However, when using the MLP classifier the classification results are improved as compared to naive Bayes Classifier. Currently, existing news documents consists of 23406 input features. But a number of selected features from the filter method are relatively higher than the hybrid method as shown in the table 7 It was quite interesting that the performance of the hybrid method didn't get worse with reduced feature subset. Consequently, a number of features reduced by filter method (Tf-idf) are relatively less so it may require another feature selection method (BPSO) for selecting an optimal feature subset.

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**

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

Table 7: **Dimension reduction**

The table 7 shows the Tf-IDF and BPSO feature selection results, how well they filtered the topics from the news documents by removing the redundant features. It is noticed that BPSO is capable of optimizing efficiently by reducing features to 7207 while Tf-IDF reduced only to 11810. The feature selection method (BPSO) was able to filter out the topics more efficiently than Tf-idf filter approach. Instead of assigning weights technique, its best to prevail with optimization based on best fit(swarm intelligence). A similar research on gene classification system proposed by Chuang et al. (2016) in which they applied Tf-idf and BPSO feature selection methods yielded better results using SVM classifier. However, it is very hard to compare the results of the proposed method with machine learning classifiers between various studies. The table 6 is about accuracy of BPSO with multi-layer perceptron and Multinomial naive Bayes classifier. However, in this study, results proved that using BPSO combined with multi-layer perceptron (MLP) classifier provides a better accuracy than Multinomial naive Bayes classifier.

6 Conclusion and Future Work

The main issue with unstructured and textual data is the lack of proper system to analyse it easily. The concept of topic selection acts as a primer in understanding the conceptual meaning behind the document(s). From a large set of document corpora, it is necessary to optimize the features and its subsets instead of just filtering based on frequency weights. The feature selection method is performed using (Tf-Idf) and BPSO. The results proved BPSO technique had better filtering process than Tf-Idf filter approach. The BPSO is combined with two different classifiers to enhance the topic selection efficiency. MLP classifier achieved an accuracy of 83%, while multinominal naive Bayes fared poor with accuracy of 80%. Usage of neural network classifier like MLP helped achieve better classification when compared to multinomial naive Bayes.

Due to time limitation, we tested the proposed model only on one set of news corpus data, in future we will extend this work and test on large corpus of data and evaluate our model performance. Since the processing time of BPSO is higher to handle high dimensional datasets this problem can be overcome using the Hadoop distributed platform. Due to its high computing capability and parallel processing of complex data which makes topic classification task easier.

References

Abualigah, L. M., Khader, A. T., Al-Betar, M. A. and Alomari, O. A. (2017). Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering, *Expert Systems with Applications* **84**: 24–36.

Akilandeswari, K. and Nasira, G. (2015). Multi-layer perceptron neural network classifier with binary particle swarm optimization based feature selection for brain-computer interfaces, *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering* **9**(6): 1615–1621.

Brito, R., Fong, S., Zhuang, Y. and Wu, Y. (2017). Generating neural networks with optimal features through particle swarm optimization, *Proceedings of the International Conference on Big Data and Internet of Thing*, ACM, pp. 96–101.

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* .

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification, *Journal of machine learning research* **3**(Mar): 1289–1305.

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* *on Robotics, Automation and Mechatronics (RAM), 2015 IEEE 7th International Conference on*, IEEE, pp. 212–217.

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.

Lang, K. (2015). 20newsgroup, *corpus database*, NewsGroup.
URL: *http://qwone.com/ jason/20Newsgroups/*

Nalluri, M. S. R., SaiSujana, T., Reddy, K. H. and Swaminathan, V. (2017). An efficient feature selection using artificial fish swarm optimization and svm classifier, *Networks & Advances in Computational Technologies (NetACT), 2017 International Conference on*, IEEE, pp. 407–411.

NLTK 3.2.5 Documentation (2015). *Documentation*, NLTK.
URL: *https://www.nltk.org/*

Pudaruth, S., Soyjaudah, K. and Gunputh, R. (2017). Markov chain carlo methods and evolutionary algorithms for automatic feature selection from legal documents, *The International Symposium on Intelligent Systems Technologies and Applications*, Springer, pp. 136–148.

Pyswarm (2015). *Documentation*, NLTK.
URL: *https://pyswarms.readthedocs.io/en/latest/*

Ratnaweera, A., Halgamuge, S. K. and Watson, H. C. (2004). Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients, *IEEE Transactions on evolutionary computation* **8**(3): 240–255.

Sahon, G. and McGill, M. (1983). Introduction to modem information retrieval, *New York: McGraw Hill* .

Souza, T. A. D., Vieira, V. J., Souza, M. A. D., Correia, S. E., Costa, S. C. and Costa, W. C. D. A. (2018). Feature selection based on binary particle swarm optimisation and neural networks for pathological voice detection, *International Journal of Bio-Inspired Computation* **11**(2): 91–101.

Ye, F. (2017). Particle swarm optimization-based automatic parameter selection for deep neural networks and its applications in large-scale and high-dimensional data, *PloS one* **12**(12): e0188746.

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.

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