International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Preprocessing Farmer Query Data Using Classic Method and Building Classifier Model

Yudhvir Singh*[1], Naresh Kumar Garg[2]
*[1]Department of Computer Science & Engineering, Maharaja Ranjit Singh Punjab Technical University Bathinda, Punjab, India
[2]Department of Computer Science & Engineering, Maharaja Ranjit Singh Punjab Technical University Bathinda, Punjab, India

ABSTRACT

Being important preliminary step, preprocessing is critical phase in text mining and other related fields. Real world data contains errors of varying magnitude with multiple interrelated issues. Data preprocessing used to transform it into a form, which is readable, acceptable by tools, data that is free from ambiguities, duplicity. In this research work, we are dealing with farmer query data set, which is kind of text data, structured in tabular form. In case of text data, before any meaningful information retrieval, preprocessing techniques are applied on the target data set to reduce the size of the data set which will increase its effectiveness. The objective of our work is to analyze the issues of preprocessing operation such as tokenization, formatting, stop word removal for our text data. After preprocessingoperations , further we have used logistic classifier to binary classify and model the farmer dataset. Logistic classifier gives good accuracy results and thus proves machine learning role in farmer query classification area.

Keywords: Machine Learning, Preprocessing, Text Data, Classification, Logistic Classifier

I. INTRODUCTION

Text data being unstructured in nature always comes with problems for mining tools. Text data is bad, in sense; it contains errors of varying magnitude with multiple interrelated issues such as duplicity, bogus information, missing values, irrelevant information. Data preprocessing is used to transform this data into a form, which is readable, acceptable by tools, which is free from ambiguities, duplicity, etc. We are dealing with farmer query dataset; it is a kind of text data, collected for state Punjab. Therefore, our major motive behind this work is to preprocess this raw farmer dataset and to gain useful information by analyzing it using text mining and machine learning tools.

There are many applications of preprocessing data in many fields. Preprocessing is a prerequisite for any useful computational output. In case of text domain, [1] discussed the role of preprocessing methods and their after effects in compression applications. Similarly, [2] used preprocessing methods in order to develop a good text to speech system. [3] used preprocessing methods in document clustering application.

After preprocessing of our farmer dataset, we are then developing a classifier model using machine learning technique based on logistic classifier. The motive is to show the stability of preprocessed dataset and to gain some insight into historical dataset using machine learning.

Processed datasets play a significant role in overall performance of any text-based mining operations. Processed data are used as input in machine learning tools to gain intelligent, inherent insight of information. Machine learning plays good role in developing intelligent systems. As for classification of web pages [4] and classification of patent databases [5] successfully applied machine learning methods.

Spam detection system developed by [6] using machine learning techniques. Classification of multisource remote sensing data using machine learning was investigated by [7]. It is also extensively being used in agricultural sector, as crop pest prediction is done by [8].

Our work is a specific attempt to discuss the preprocessing possibilities in case of agriculture dataset domain, which is unique, and no such similar attempt exists to best of our knowledge.

II. DATA DESCRIPTION

Text dataset that we are using collected from [9]. These are agriculture sector dataset related to farmer queries. Our geographical area of interest is Punjab state, India for which data is collected for years, 2015 and 2016. Our research work deals with these raw datasets and focuses on preprocessing of these datasets.

III. PREPROCESSING METHOD

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

A. Software's and Tools Used

In this research work, combination of software's and tools are used to perform various tasks. For basic computational task we have used MS package, various online, offline formatting tools. For text mining and other learning tasks, we have used python and libraries scikit-learn,graphlab, pandas [10], sframe, nltk and other related packages.

B. Data Preprocessing Workflow

Text data being unstructured in nature always comes with multitude of problems for various kind of text mining tool. To get some useful insights from this raw data, foremost important step is to preprocess it. Since in our case, collected farmer dataset is also comes with errors, issues, so preprocessing is performed to make it clean and input worthy. Using bag of words (BOW) concept, tokenization of text is done. Individual word acts as a token unit to generate feature vectors. For preprocessing we have adopted classic method which usage a prebuilt stop words list to remove most frequent words and bogus information [11]. Similarly, other operations are performed like uniform formatting, punctuation removal, etc. Following Figure 1 shows work flow of the adopted methodology.

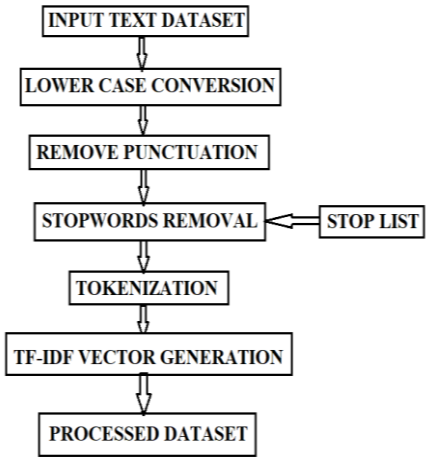

Figure 1. Preprocessing workflow

C. Data Unification

Segregated raw data for individual districts needed to be combined so that a unified state level sframe data file can be generated which will be more meaningful to our approaches for preprocessing and other mining operations, rather than working on each individual sframe separately and then combining the results. Following Fig. 2 shows the whole process.

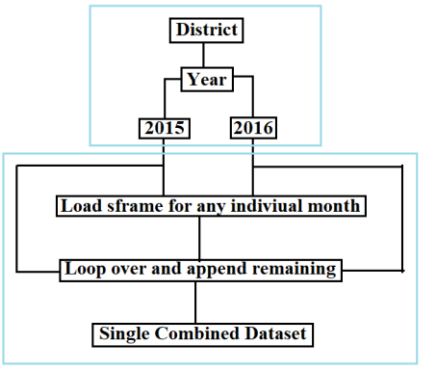

Figure 2. Data unification process.

D. Weighting Vectors

Term frequency - inverse document frequency (TF-IDF) is used as a weighting criterion for feature vectors for our farmer text dataset. We are using TF-IDF for present research work because for algorithmic analysis text data needed to be represented in numerical form and TF-IDF score is the popular criteria in text mining field. TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection [12].

IV. DATA CLASSIFICATION

Preprocessed dataset is used as input in machine learning tools to get some intelligence about farmer dataset. Our dataset comprising of two measure categories which was used as class labels for classification problem. We have performed binary classification of our dataset and developed machine learning model.

TF-IDF score generated for our data set. Then using this vector as a basic numerical representation, supervised classification performed.

related packages.

A. Algorithm Used

We have used logistic classifier (LC) to model farmer dataset. In LC the dependent variable (DV) is categorical. In case of a binary dependent variable, that is, where the output can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick.

Following general algorithmic steps are followed:

- Input Farmer Dataset (FD)
- Divide FD into Training(train_data) and Test (test_data) Data
- Select a logistic classifier as machine learning algorithm
- Use train_data for building classifier model
- Use test_data for model prediction
- Output result

B. Evaluation Metrics

In this paper, confusion matrix is generated for classification result and accuracy, misclassification rate and receiver operating characteristic (ROC) are used to evaluate algorithmic performance. Following are the formulas for these metrics:

$$\text{Accuracy} = (TP+TN) / (P+N) \quad (1)$$

$$\text{Misclassification rate} = (FP+FN) / (P+N) \quad (2)$$

TP, TN, FP, P, N refer to the number of true positive, true negative, false positive, positive and negative samples, respectively.

V. RESULT AND DISCUSSION

A. Preprocessing Result

Output of preprocessing operations is data, which was free from ambiguities, errors, null values, duplicities or any other unnecessary symbols and bogus information. Following Figure 3 is the sample file comparison of raw file before preprocessing and preprocessed, clean file.

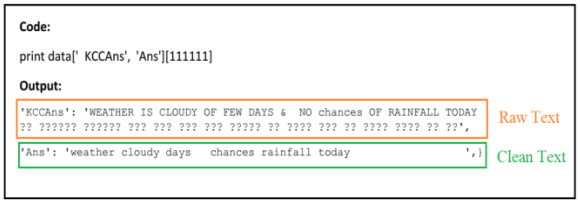

Figure 3. Comparison of raw and processed sample data.

B. Classification Result

Confusion matrix for binary classification result using logistic classifier is shown below in Table 1.

Table 1. Confusion matrix

		Predicted Label	
		TP	FP
Target Label	TP	11037	1505
	FP	1257	9046

Following Table 2 gives the performance statistics of binary classification for logistic classifier.

Table 2. Performance Statistics

Algorithm	Accuracy	Misclassification Rate
LC	87.86%	12.14%

As given in above Table 2 logistic classifier predicting classes with accuracy 87.86% or 12.14% misclassification rate. So, in clear terms LC algorithm is performing well on our farmer dataset. The accuracy, and misclassification rate of LC algorithm is shown below in Figure 4.

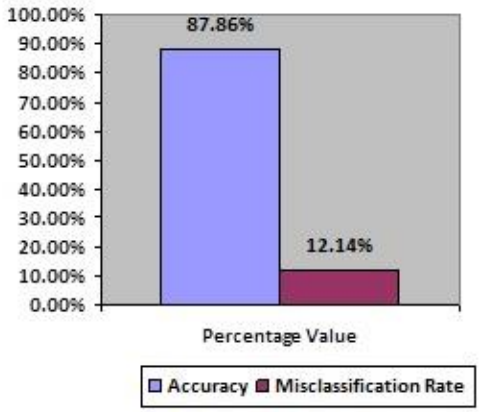

Figure 4. Performance metrics.

Following Figure 5 shows the ROC curve of our classification result.

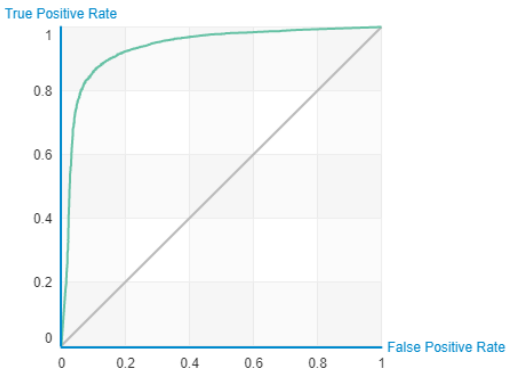

Figure 5.ROC evaluation.

So, based on above roc curve, we can state that logistic classifier model is performing well on preprocessed farmer dataset.

VI. CONCLUSION

This research work deals with preprocessing work on farmer query dataset. We used classic preprocessing technique, available for mining text data, and performed operations such as stop words elimination, formatting, punctuation removal. This work gives some insight about text mining and tries to provide good knowledge about various preprocessing techniques available for text mining. A small experimental introduction to the modern tools, software's, machine learning libraries being used for preprocessing text data is given, so that it can somewhat overviews all those approaches that can be adopted with text data.

Starting with raw text data related to farmer queries and responses, we worked to make it loadable, readable by removing all the errors and overcoming major issues. Then we performed data reduction operation and reduced huge dataset into unified singe datafile. After unification of data, preprocessing operations are performed using python language and supported libraries. After all these steps, finally we come up with a processed dataset which is ready to be given as input to various data mining and machine learning tools and to be analyzed. At the end, we performed binary classification of farmer dataset using logistic classifier to show that data is stable and

can be used for mining tasks. Logistic classifier performs well in our case and gives around 87.86% of accuracy. Therefore, we can say that data preprocessing is very essential phase as without it we cannot proceed in a manner that make our work useful, interpretable and result oriented.

VII. REFERENCES

[1]. J. Abel and W. Teahan, "Universal text preprocessing for data compression," IEEE Transactions on Computers, vol. 54, no. 5, pp. 497-507, May 2005. DOI: 10.1109/TC.2005.85

[2]. H. Jing, "Identifying accents in Italian text: a preprocessing step in TTS," IEEE Workshop on Speech Synthesis, pp. 151-154, 2002. DOI: 10.1109/WSS.2002.1224396

[3]. A. I. Kadhim, Y. N. Cheah and N. H. Ahamed, "Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering," 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, Kota Kinabalu, 2014, pp. 69-73.DOI: 10.1109/ICAIET.2014.21

[4]. W. T. Aung and K. H. M. S. Hla, "Random forest classifier for multi-category classification of web pages," IEEE Asia-Pacific Services Computing, 2009, pp. 372–376. DOI: 10.1109/APSCC.2009.5394100

[5]. F. Al Shamsi and Z. Aung, "Automatic patent classification by a three-phase model with document frequency matrix and boosted tree,"International Conference on Electronic Devices, Systems, and Applications, 2017. DOI: 10.1109/ICEDSA.2016.7818566

[6]. D. K. Renuka, T. Hamsapriya, M. R. Chakkaravarthi and P. L. Surya, "Spam Classification Based on Supervised Learning Using Machine Learning Techniques," International Conference on Process Automation, Control and Computing, Coimbatore, 2011, pp. 1-7. DOI: 10.1109/PACC.2011.5979035

[7]. P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random forest classification of multisource remote sensing and geographic data,"IEEE Int. Geosci. Remote Sens. Symp. (IGARSS 2004), vol. 0, no. C, pp. 1049–1052, 2004.

[8]. Y. H. Kim, S. J. Yoo, Y. H. Gu, J. H. Lim, D. Han, and S. W. Baik, "Crop Pests Prediction Method Using Regression and Machine Learning Technology: Survey,"IERI Procedia, vol. 6, pp. 52–56, 2014.

[9]. "Open Government Data (OGD) Platform India."Online]. Available at: https://data.gov.in/. Accessed 11 June 2017].

[10]. "Python Data Analysis Library."online] Available at: https://pandas.pydata.org. Accessed 19 Aug. 2017].

[11]. Anjali, et al "A comparative study of stemming algorithms". Int. J. Comp. Tech. Applications. Vol. 2, pp:1930-1938, 2007.

[12]. J. Leskovec J, A. Rajaraman A, J.D. Ullman, Data Mining. In: Mining of massive datasets, 2nd edn. Cambridge University Press, Cambridge, 2014, pp 1-18.