© 2018 IJSRSET | Volume 4 | Issue 4 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Text Classification using Data Mining and Machine Learning Techniques: A Brief Review

Diksha Kose, Mayuri Harankhede, Shivani Shukla, Prof S.W.Mohod
Department of Computer Engineering Bapurao Deshmukh College of Engineering Wardha, Maharashtra, India

ABSTRACT

Text classification has grown into more significant in managing and organizing the text data due to tremendous growth of online information. It does classification of documents in to fixed number of predefined categories. Rule based approach and Machine learning approach are the two ways of text classification. In rule based approach, classification of documents is done based on manually defined rules. In Machine learning based approach, classification rules or classifier are defined automatically using example documents. It has higher recall and quick process. This paper shows an investigation on text classification utilizing different machine learning techniques.

Keywords: Text Classification, Machine Learning, Multiview Boosting, cMFDR, Visual Classifiers, Text Segmentation

I. INTRODUCTION

Text classification is the process of separating a bunch of text records in to various categories in the defined order. It will so sorting and arranging the records in the folders hierarchies so that identification of topic become easy that would support topic specific operation. But if this work done by manually means, it‟s a time consuming work, error phone and needs expert of domain in predefined category. It is necessary to use machine learning techniques automate the classification work. The text classification process has many steps. Collect all format documents and do preprocessing. It is done by token and stem word generation. Removing the insignificant words by token generation and apply the stemmey procedure used for convert the different word format to canonical formal. Clear format is created at the final stage of preprocessing.

Machine learning is band on algorithm that will define classification rules or classifiers automatically using example documents. Two types of learning algorithm is used in the machine learning-supervised learning algorithm or unsupervised learning algorithm. Earlier days, Machine learning was used of binary classifications the classifier rules determines whether the record belongs to defined set or not. For eg. Using machine learning, in email/binary classification is done by whether the mail belongs to legitimate or spam. Similarly machine learning is applied to multi label classification. Here we analyzed seven different machine learning techniques after text classifications by different authors for many applications. In order to handle it in easy way apply indexing to the record collection and create record vector. In order to create vector space, feature selection is the vital step of text classification and utilized for selection of features subset from original records Finally classification of records is prepared by either manually or automatically.

II. AUTOMATIC FEATURE SUBSETS ANALYSER

Developed by Rogerio C. P. Fragoso et al.

Basic idea: The increasing accessibility of text documents in digital form is leading to a need for easy, flexible and automated ways of accessing their contents. In this context, text categorization (TC) is a crucial tool for content organization and information retrieval. The AFSA enhances the already present Class-dependent Maximum Features per Document (cMFDR) procedure and spontaneously describes the good count of features per document. In the cMFDR procedure, the feature's count is chosen after a concurrent application of the methods which is a long time-taken strategy. In contrast, AFSA finds the good count of features in a data related driven way which is quicker than cMFDR.

Outline of concept:

Many of the text categorization techniques use the Bag of Words for representation. In this method, every word of a textual document is considered a feature. Thereby, a standard sized dataset should have thousands of features, which makes text categorization a great dimensional problem. Calculation costs of categorization are directly affected by the dataset's dimensionality. Furthermore, the excessive features can negatively impact on the classification accuracy, especially in datasets having a minimal number of instances. As many of the features in text documents are redundant or irrelevant for classification, these issues can be addressed by restricting the feature's count in the dataset, this method to be called as dimensionality reduction. Dropping the feature's count in the dataset can improve computational efficiency while managing or enhancing classification performance. Dimensionality reduction techniques are feature extraction and feature selection. The first technique based on extracting approach that dynamically determines the best value for the f parameter, i.e., the count of features to be chosen per document. The cMFDR method presents good results, however, as its predecessors, it requires a value for the parameter f. In a production environment, as new documents are presented, it is necessary, from time to time, to reanalyze the dataset including the new documents in the feature selection process. Every time this analysis is performed, the client needs to test manually the subsets generated by cMFDR and choose the best one. So, in a working environment, different values of the parameter "f" must be evaluated to determine which one generates the best subset. This process is time-consuming and requires human interaction. Automatic Feature Subsets Analyzer (AFSA), the proposed method, aims to establish, in a data-driven way, the best value for f parameter and, hence, the feature"s count is taken. The proposed method adopts a validation set strategy to generate a sequence of subsets using cMFDR, reassesses the performance of each set and then choose the value for f that produces the effective set. So, the best value off is provided as input for cMFDR to build the final feature set, using the test set.

III. MINING USING LATENT DIRICHLET ALLOCATION (LDA)

Developed by Carina Silberer et al.

Basic Idea: LDA based topic place and to filter the high level representation in mapped text reports. But this method is highly sensitive to high hierarchy parameters related to the number of classes or topics and there is no processed way to forecast correct configuration. In the same manner, the relevant variability yields from the content common to the document and the content of each classes composing the topic place. The proposed -vector representation is evaluated in the context of the theme identification of automatic transcription of human-human telephone conversations and of the categorization of textual newswire collection. This representation is built from a set of feature vectors. Each one is composed of scores of discriminative words. Then, the metric used to associate a document to a class is the Mahalanobis metric. Outline of concept: The first step is to build a set of topic places from a training corpus of dialogues D={d1,d2,…dn}. LDA algorithm with different high hierarchy parameters used to learn these topic places. Then, each dialogue is mapped into

each topic place to obtain a set of topic-based representations of the same document. This representation is heterogeneous. Indeed, the feature space is not the same from one LDA hyper-parameter configuration to another. To tackle this issue, we propose to map each topic-based representation into a common discriminative set of words. Thus, we obtain a set of homogeneous feature vectors to represent each document, and each feature is the probability that a discriminative word is associated with a given document. Then, these multiple views are compacted with the -vector framework. This process allows us to remove most of the useless and irrelevant information from the topic-based presentations of the document. In this research, deal various representation of reports or speech transcriptions and fusion process with factor analysis process called i-vector. The first step is to present a report in multiple/various topic places of different sizes. Then a compact representation of the report from is used to compensate the vocabulary and variability of the topic named presentation. This method was developed for speaker new algorithm. Then it is applied to text categorization task. All original topic band presentation of documents (c-vector) is joined with EFR normalization also provides good solution to report variability's. This method results 86.5% accuracy in classification compared to other LDA Speaker.

IV. SCENE TEXT DETECTION AND SEGMENTATION

Developed by Youbao Tang, and Xiangqian Wu

Basic idea: The objective of scene text detection is to locate the positions of text in different scenes, for example guideposts, store marks, and warning signs; it is one of the most important steps for end-to-end scene text recognition. It can enhance the performance of various multimedia applications, e.g. mobile visual searches, content-based image retrieval, and automatic sign translation.VGGNet-16 is modified for framing the framing the CNN based models. The boundaries and the entire region of text is used for training and designing of a detection network called DNet which is a CNN based textaware candidate text region used for detecting the coarse CTR. For segmenting the detected coarse CTR into the refined CTRs, SNet a segmentation network which is based on CTR refinement model has been applied. When compared with the traditional approaches, these SNet and DNet networks used for extracting very few CTRs so more true text regions are kept. To get the final text region from refined CTRs, CTR classification network CNet is defined. Most existing scene text detection approaches generate text bounding boxes containing a lot of background, which makes scene text recognition difficult. To deal with this issue, scene text segmentation methods were proposed to obtain more precise text regions. These methods were hand-crafted and cannot be trained by an end-to-end process. To solve these problems, in this method, a text-aware CTR extraction model and a CTR refinement model are devised to extract CTRs and obtain precise text segmentation results, which can overcome the above problems.

Outline of concept:

In this model, first give the input image and CNN model to predict a text-aware salient map. The text region having a larger values is fusion probability map which is used as the final salient map of the input image. Then extract the coarse CTR to get an accurate saliency prediction, the CNN architecture should be deep and have multi-scale stages with different strides. Finally, the refined CTRs are fed into a CTR classification model to filter out non-text regions and obtain the final text regions. Next step Refinement, The extracted coarse CTRs usually contain some background regions. Because of the diversity of texts and backgrounds in scene images, it is impossible to consider all cases in the training dataset. It should contains some errors in the coarse CTRs. And when the texts are close to each other, multiple words or text lines will be considered as one text region in the coarse CTRs. the text detection result will reduce

recall and precision. Moreover, accurate text segmentation can provide helpful information for scene text recognition. Based on the coarse CTRs, it is necessary to further refine the text region segmentation results. Some non-text regions occur in the CTR refinement results. Therefore, given a CTR image, there is a need to classify it into text or non-text, which is actually a two-class problem in image classification. After filtering out the non-text regions with CNet, cluster the text regions into text lines according to their vertical locations and heights. Then the text lines are separated into words as the final text detection results according to the distances of adjacent text regions in the same text line. Therefore, the proposed text-aware CTR extraction model can extract more true text regions and much fewer false text regions than other approaches.

V. CONCLUSION

We have been studied the different machine learning techniques executed for text classification. This investigation gave a review of probably the most principal design and procedures which are broadly utilized in the text area. Each methodology have their own remark in their applications.Boost.SH algorithms performs better than the other procedure implemented in multi view learning applications. AFSA achieves better results than existing cMFDR with less training data. Inconsistency detection can yields high performance when done using multilevel text categorization followed by conceptual reasoning. Scene text detection method by using multiple convolutional neural networks provides comparable precision on the bench mark datasets. The proposed kernelized version generated by multiple kernel learning provides better results. The model using stacked auto encoders provides better fit to behavioral data compared to baselines.

VI. REFERENCES

[1]. C.L.Liu, W.H.Hsaio, C.H.Lee, T.H.Chang, T.H.Kuo, "Semi-Supervised Text Classification with Universum Learning", IEEE Trans. on Cybernetics, Vol.46, Issue.2, P.462- 473, 2016.

[2]. J.Peng, A.J.Aved, G.Seetharaman, K.Palaniappan, "Multiview Boosting With Information Propagation for Classification", IEEE Trans. on Neural Networks and Learning Systems, 2017.

[3]. F.Zhuang, P.Luo, C.Du, Q.He, Z.Shi, H.Xiong, "Triplex Transfer Learning: Exploiting Both Shared and Distinct Concepts for Text Classification" IEEE Trans. on Cybernetics, Vol.44, Issue.7, PP.1191-1203, 2014.

[4]. R.C.P.Fragoso, R.H.W.Pinheiro, G.D.C.Cavalcanti, "A method for automatic determination of the feature vector size for text categorization", 5th Brazilian Con. on Intelligent Systems, 2016.

[5]. J.A.Otaibi, Z.Safi, A.Hassaine, F.Islam and A.jaoua, "Machine Learning and Conceptual Reasoning for Inconsistency Detection", IEEE Access, 2017, Vol.5, PP.338-346.

[6]. S.Baccianella, A.Esuli, F.Sebastiani, "Feature Selection for Ordinal Text Classification", Neural Computation, Vol.26, Issue.3, Pages.557-591, 2014.

[7]. B.Zhang, A.Marin, B.Hutchinson, M.Ostendorf, "Learning Phrase Patterns for Text Classification", IEEE Transactions on Audio, Speech, and Language Processing, Vol.21, Issue.6, PP.1180-1189, 2013.

[8]. J.Y.Jiang, R.J.Liou, S.J.Lee, "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification", IEEE Transactions on Knowledge and Data Engineering, Vol.23, Issue.3, PP.335 - 349, 2011.

[9]. C.Silva, U.Lotric, B.Ribeiro, A.Dobnikar, "Distributed Text Classification With an Ensemble Kernel-Based Learning Approach", IEEE Transactions on Systems, Man, and

Cybernetics, Part C (Applications and Reviews) Vol.40, Issue.3, PP.287 - 297, 2010.

[10]. Y.Tang, & X.Wu, "Scene Text Detection and Segmentation based on Cascaded Convolution Neural Networks", IEEE Transactions on Image Processing, 2017, Vol.26, Issue.3, PP.1509 - 1520.

[11]. M.Elhoseiny, A.Elgammal & B.Saleh, "Write a Classifier: Predicting Visual Classifiers from Unstructured Text", IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, Vol.39, Issue.12, PP.2539 - 2553.