International Journal of Recent Trends in Engineering & Research
ISSN (ONLINE) : 2455 – 1457
IMPACT FACTOR : 4.101

CHALLENGES IN TEXT CLASSIFICATION USING MACHINE LEARNING TECHNIQUES

Jaspreet Singh
Research Scholar, Department of Computer Science, Guru Nanak Dev University, Amritsar (Punjab), India

Abstract: The era of machine learning has turned every unturned stone by enlightening enormous range of applications through the introduction of machine learning. Text classification is one such area from the world of information retrieval where real world applications leveraged machine learning classifiers. Nevertheless, the traditional and state-of-the-art machine learning classifiers did wonders but, the introduction of neural network processing to text classification has broken many dependencies of classical methods. Recurrent neural networks outperforms in the field of text classification than Convolution neural networks which did wonders in computer vision field. This article highlights some challenges of text classification process using machine learning techniques.

Keywords: Text Classification, Machine Learning Techniques, Feature engineering, Challenges

I. INTRODUCTION

Text classification involves the categorization of text into its corresponding class thereby discriminating other classes in the data. The health of text classifier is taken as good one if it has gained a prior knowledge of all the cases with their respective classes. The main concern is training data; sampling techniques must be appropriately applied in order to get correct training of classifier. In many cases, the downfall in performance of text classification is observed due to misinterpretation of class of text into some other similar class. The challenge of correct classification is significantly taken by deep learning introduction into natural language processing. However, few challenges are still alive in the course of machine learning applications to text classification processes.

II. CHALLENGES IN TEXT CLASSIFICATION

The following are few challenges observed in text classification through machine learning.

A. Feature vectors must acquire complex semantics of text.
B. Binary or numeric features obtained from word of phrase frequency must be noise free.
C. The structure of classifier like Naïve Bayes uses high dominance of model rather than hidden text features thereby suppressing performance of classifier.
D. Information retrieval systems experience diverse nature of texts with highly variable content, quality and length.
E. The performance of machine learning model would be degraded if an ill-sampled data is presented to it while training and some classes are not observed by it.
F. During the training phase of machine learning model, the attained knowledge often escaped from given real data and hence resulting in deterioration of performance.
G. The classification of text sometimes goes more subjective due to presence of unknown classes and outliers.

H. The volume of training data plays a huge role in learning a model. Training data must be labeled and large enough to cover all the upcoming classes.
I. Human labelers expressively bias the training data which may yield a wrong training of model.
J. Sometimes a text classification problem consists of large number of closely related classes, for example: Google directory contains around two billion categories in a deep hierarchy hence, making it difficult to correctly classify the test data through machine learning.
K. According to the scope of text, the number of text features may vary from few hundreds to few thousands of thousand features. A good choice of feature engineering can significantly attain a great deal of mileage in text classification process.
L. In case of large volumes of training data, stemming and lower-casing may decline the performance of statistical machine learning techniques. For example: words like 'oxygenate' and 'oxygenation' yields 'oxygen' as an outcome of stemming thereby thrashing the real semantics of text.

III. RELATED WORK

Challenge	Reference	Machine Learning Classifier
A	[1], [2]	Support Vector Machine, Naïve Bayes, and J48
B	[1], [3], [4]	Ensemble Classifier
C	[3]	K-nearest Neighbor
D	[2], [4], [5]	Naïve Bayes
E	[4], [6]	Latent Dirichlet Allocation
F	[6], [13], [8]	Decision Tree, Random Forest
G	[5], [9]	Naïve Bayes
H	[3], [10], [11]	Naïve Bayes
I	[4], [7], [12]	Support Vector Machine
J	[8], [9], [14]	Decision Tree, Logistic Regression
K	[4], [5]	Naïve Bayes
L	[3], [11]	Support Vector Machines

Table 1: List of challenges along with their references and classifiers

IV. CONCLUSION

This article threw light on various challenges that machine learning classifiers are facing during the task of text classification. One of the main solutions to limited training data is bootstrapping process, which involves initial training of model with few initially labeled data points, further unlabeled data in bulk will be supplied which can attempt to perform further learning of classifier. Another challenge of labeling the huge training data can be resolved by exploiting wide range of labelers in order to reduce biasness. The issue of feature engineering can be lifted up by deploying separate machine learning for feature extraction process in the pre-processing task of text classification.

REFERENCES

I. Banerjee, Somnath. "Boosting inductive transfer for text classification using wikipedia." Sixth International Conference on Machine Learning and Applications (ICMLA 2007), 2007.
II. Burkhardt, Sophie, and Stefan Kramer. "Online multi-label dependency topic models for text classification." Machine Learning, 2017.

@IJRTER-2018, All Rights Reserved

III. Chirawichitchai, Nivet. "Emotion classification of Thai text based using term weighting and machine learning techniques." 2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2014.
IV. "Classification Learning." Encyclopedia of Machine Learning and Data Mining, 2017, pp. 209-209.
V. Holts, Alberto, et al. "Automated Text Binary Classification Using Machine Learning Approach." 2010 XXIX International Conference of the Chilean Computer Science Society, 2010.
VI. Jiang, Eric P. "Content-Based Spam Email Classification using Machine-Learning Algorithms." Text Mining, 2010, pp. 37-56.
VII. Joachims, Thorsten. "Text Classification." Learning to Classify Text Using Support Vector Machines, 2002, pp. 7-33.
VIII. Lewis, David D. "Challenges in machine learning for text classification." Proceedings of the ninth annual conference on Computational learning theory - COLT '96, 1996.
IX. Mohasseb, Alaa, et al. "Domain specific syntax based approach for text classification in machine learning context." 2017 International Conference on Machine Learning and Cybernetics (ICMLC), 2017.
X. Rouigueb, A., et al. "Text-independent MFCCs vectors classification improvement using local ICA." 2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2013.
XI. Salur, Mehmet U., et al. "Text classification on mahout with Naïve-Bayes machine learning algorithm." 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), 2017.
XII. Sharma, Neeraj, et al. "Text Classification Using Hierarchical Sparse Representation Classifiers." 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017.
XIII. Trindade, Luis A., et al. "Text classification using word sequence kernel methods." 2011 International Conference on Machine Learning and Cybernetics, 2011.
XIV. Wong, et al. "Using complex linguistic features in context-sensitive text classification techniques." 2005 International Conference on Machine Learning and Cybernetics, 2005.

@IJRTER-2018, All Rights Reserved