IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Linear Support Vector Machine (SVM) with Stochastic Gradient Descent (SGD) training & multinomial Naïve Bayes (NB) in News Classification

Feroz Ahmed[1], Shabina Ghafir[2*]

[1, 2] Department of Computer Science & Engineering, Jamia Hamdard, New Delhi, India

[*]*Corresponding Author: sghafir@jamiahamdard.ac.in, Mob.:9716228928*

DOI: https://doi.org/10.26438/ijcse/v7i4.360363 | Available online at: www.ijcseonline.org

Accepted: 13/Apr/2019, Published: 30/Apr/2019

***Abstract*—** The motivation for this work arises from the need of Automatic Document Classification (ADC) which is necessary when the task involves business specific contexts which cannot be fulfilled via querying on any search engine. Since a large number of websites are available over the internet nowadays therefore users generally search information from different websites via search engines now. But search engines require appropriate keywords from users in order to give relevant information from the web and sometimes users have obtained irrelevant results if he or she is not sound enough to provide keywords correctly. Thus, we need a proper document classification for the material of our wish so that the one which is required can be obtained easily instead of wasting time in searching. To understand this, we have discussed the automatic document classification in news domain where we classify news articles into four distinct categories: business, science & technology, entertainment and health using Linear SVM with SGD training and multinomial NB classifier and compare their performance. The classification is based on the title of the news article taken as feature.

***Keywords*—**Automatic Document Classification, Linear SVM, Stochastic Gradient Descent, multinomial NB

I. INTRODUCTION

Automatic document classification [1] is the task of assigning predefined categories to text documents so that they can be easily retrieved while searching. As the size of online information increasing exponentially, this task is of great significance. Even searching information via search engines will not yield totally satisfactory result unless users do not enter the exact keywords in search engines to get the relevant information.

Also, simply searching will be a waste of time and in this age of digital world we cannot afford manual classification of documents over the internet. Thus, in place of manual classification of documents, we have proposed in this work that machine learning algorithms can be trained to classify documents based on human-labelled training documents. In this work, we have discussed classification of news pages into five distinct categories i.e. business, science & technology, entertainment and health using Linear SVM with SGD training and multinomial NB classifier which is often used as a starting point in text classification. The news page classification technique uses a variety of information to classify a target page. The essential concept regarding classification of news pages are the attributes of news pages. In this work, we have taken title of the news article as our feature variable and category of the news article which we will predict using the classifiers as our target variable. A news page title is the cheapest to achieve and significant source for classification. The classification technique uses this feature to segregate news pages into different categories.

The rest of the work is described as follows. Section 2 consists of the literature review. Section 3 discusses the proposed viewpoint regarding news page classification in detail. The details of the dataset used in our work have been discussed in section 4 along with the analysis of the result so obtained. At the last, in section 5 we have concluded our work.

II. RELATED WORK

A vast amount of work has been done in the field of text classification related to news. Taeho Jo [2] proposed a modified version of K Nearest Neighbor (KNN) for text classification. He defined the similarity measure which was based on the semantic relations among words to consider both attribute and attribute values between representations of texts. Using the similarity measure he modified the traditional KNN. Shuo Xu, Yan Li and Zheng Wang [3] proposed a Bayesian version Naïve Bayes (NB) classifier for text classification. They applied that Bayesian version on 20 newsgroup dataset with appropriate Dirichlet hyper-parameters. Abdelaali Hassaine, Souad Mecheter and Ali

Jaoua [4] represented a corpus of documents by a binary relation linking each document to the word it contains. They made use of the Hyper Rectangular Algorithm to extract the list of the most representative words in a hierarchical way from the relation being made. They then fed the extracted keywords into the random forest classifier in order to foretell the category of each document. Ari Arulia Hakim, Alva Erwin, et al. [5] performed the classification of news articles in Bahasa Indonesia using Term Frequency Inverse Document Frequency (TF – IDF) algorithm. Even, the concept of deep learning has been applied in news classification. Shuang Qiu, Mingyang Jiang, et al. [6] proposed a new kind of Stacked Denoising Auto Encoder (SDAE) algorithm which they termed as LMSDAE algorithm. They applied this newly version for classification of Chinese news articles and found it better than other three algorithms – SDAE, Sparse Denoising Auto Encoder (SPDAE) and Deep Belief Nets (DBN) respectively because of the reduced training times and increased convergence rate. Similarly, Chenbin Li, Guohua Zhan, et al. purposed an improved Bi-LSTM-CNN [7] in news text classification. Also, the classification of Chinese text has also been discussed in [8] using semantic kernel in SVM. Ankit Dilip Patel and Yogesh Kumar Sharma [9] proposed Web Page Classification (WPC) on newsfeeds to recognize and allocate them into categories of news like sports, business, world, health in order to enhance user's accessibility towards relevant news by pass over suitable category as per user's choice. This has been done by choosing hybrid technique of URL analysis and content context analysis. Fasihul Kabir, Sabbir Siddique, et al. [10] performed Bangla text document categorization using SGD classifier. They performed their experiment on BDNews24 documents.

Thus, after exploring the latest work discussed above in news classification, we observed that majority of the work has been done in mainly Naïve Bayes, SVM, Neural Networks and k-NN which has also been mentioned in [1] and [11] besides few exceptions. Hence, in this work we decided to explore further by classifying news articles using a linear SVM with SGD training and compare its performance with multinomial Naïve Bayes.

III. METHODOLOGY

For the classification of news articles correctly, we have chosen title of the news article as an attribute on the basis of which we classify a news page as business, science & technology, entertainment or sports article. Then linear SVM with SGD training and multinomial NB are used to classify news article. We evaluate the performance of both the classifiers by using Spyder tool.

The details of the system environment for this research are:

a. Processor: Intel(R) Core(TM) i5-4210U CPU @ 1.70GHz
b. Hard Disk Drive: 1 TB
c. Installed Memory: 4 GB
d. Operating System: Windows 7 64 bit
e. Developed under Python Programming Language.

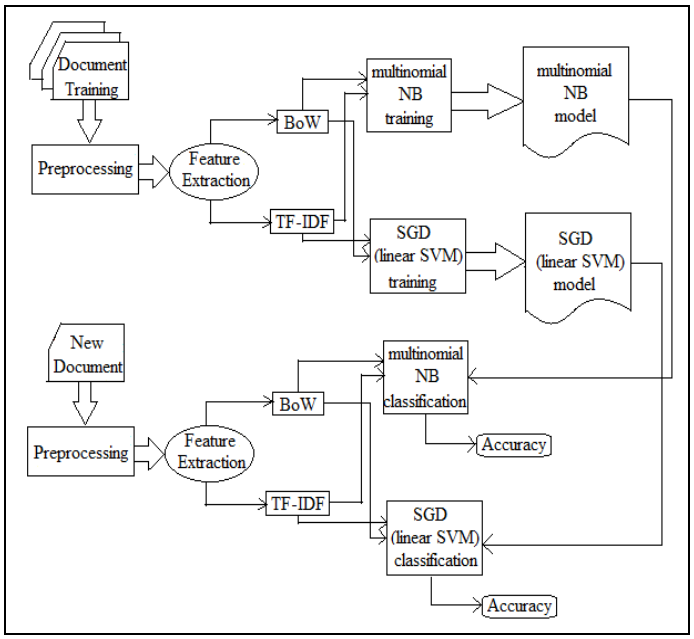

Fig. 1. Proposed Diagram for ADC

The diagram in figure 1 represents the proposed ADC methodology in this work where the process has been divided into training and testing process respectively.

A. *Preprocessing of dataset*

In order for better results, preprocessing of data must be done. We decided to transform categories into discrete numerical values followed by transformation of all English alphabet letters into lowercase so that the dataset become consistent. Besides these, removal of punctuations is also necessary to reduce the processing time. Hence we removed all the punctuations from our dataset.

B. *Feature Extraction*

Although feature extraction is desirable in classification tasks, it is having utmost importance in ADC because of the high dimensionality of text features and the existence of noisy features.In feature extraction from the dataset we have utilized the following two approaches with both the SGD (linear SVM) and multinomial NB classifiers.

1) *Bag of Words (BoW)*

In BoW [1], a document is represented as a set of words, combined with their associated frequency in the document. Such a representation is essentially independent of the sequence of words in the corpus.

2) TF-IDF

TF-IDF or term frequency – inverse document frequency. It make sure that a term is an important indexing term for a document if it occurs frequently in it while on the contrary terms which occur in many documents are rated less important indexing terms because of their low inverse document frequency.

$$W_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i})$$

$W_{i,j}$ = TF-IDF weight of a term computation result.

$tf_{i,j}$ = frequency of ' *i* ' in ' *j* ' i.e. how many times a term can be found in a text or a category.

$\log(\frac{N}{df_i})$ = computation of IDF where *'df_i'* is number of documents containing term and *'N'* is total number of documents.

C. Learning Algorithms

1) SGD Classifier

Gradient descent is one of the mostly used algorithms that can offer new outlook for solving problems. It is an algorithm to minimize functions. When given a function that defined by a set of parameters, it begins with an initial set of parameter values and makes iteration to move toward set of parameter values that find minimal point for the function. This minimization process utilizes derivation in calculus to find aligned line that approaching the minima. Gradient descent can't be fast enough to run on very large datasets. One iteration of this algorithm requires a prediction of each instance in the training dataset, it may take a long time when have millions of instances. Stochastic gradient descent is slightly different because the coefficient update occurs only when training process is running. The update methodology for the coefficient is the same as Gradient Descent, except the cost is not summed over all training patterns, instead only calculated for one training pattern.

SGD is an efficient technique to discriminative learning of linear classifiers under convex loss functions. In our case, we have utilized SVM as a loss function by keeping loss parameter equals to 'hinge', l2 regularization as penalty, keeping value of alpha equals to 0.0001, number of iterations equals to 5 and random_state equals to 42. Thus, the resultant classifier will be linear SVM with SGD training.

2) Multinomial NB Classifier

The job of this multinomial model is to find the frequency of different terms used in any document by representing a document with a BoW. Then, the documents in every class can be modelled as samples drawn from a multinomial word distribution. Consequently, the conditional probability of a document given a class is just a product of the probability of each observed word in the corresponding class.

It is appropriate for the purpose of classification having discrete features which are based on integer counts however fractional count for example TF-IDF also helpful. In our case, we have utilized it with both BoW and TF-IDF features by keeping additive smoothing parameter to 1.0 i.e. alpha=1.0, fit_prior=True and prior probabilities of the classes to none i.e. class_prior=None thus maintaining all parameters to their default behaviors.

IV. RESULTS AND DISCUSSION

The dataset used in this experiment has been obtained from kaggle [12]. The dataset consist of headlines, URLs and categories for 422419 news stories. The dataset contain news related to business, science & technology, entertainment and health categories. We split our dataset in a training sequence that contains 316814 news stories and a test sequence comprises of 105605 news stories. We have made use of the Spyder tool for performing our experiment. We performed our experiment with both multinomial NB and SGD classifier (linear SVM) respectively, each utilizing both BoW and TF-IDF in feature extraction. The results obtained after performing the experiment are shown below in the table.

Table 1 Experimental Results (in Percentage)

		Comparison of results			
		Accuracy	Recall	Precision	F1
multi-nomial NB	BoW	92.83	92.83	92.83	92.83
	TF-IDF	92.55	92.55	92.60	92.53
SGD (linear SVM)	BoW	89.25	89.25	89.34	89.12
	TF-IDF	84.26	84.26	85.35	83.93

Thus, we can see that the multinomial NB classifier with BoW in feature extraction outperforms all other techniques by maintaining an accuracy rate of 92.83% while having a close margin with TF-IDF i.e. 92.55%. However, although SGD couldn't manage to qualify as more accurate but it is not too far in accuracy from multinomial NB when combined with BoW approach; a round off of 3% in difference but in case of its combination with TF-IDF it lags behind from multinomial NB by approx. 10%.

V. CONCLUSION AND FUTURE SCOPE

This work is aimed at performing ADC using machine learning algorithms. With the advent of massive online text, the role of ADC has become prominent. In our work, we explore ADC in the domain of news where we compared the performance of multinomial NB and SGD (linear SVM) and found multinomial NB better than the other.

ACKNOWLEDGMENT

I express my gratitude to my research guide, Ms. Shabina Ghafir (Asst. Prof.) for her guidance, co-operation and support to carry out this research work.

REFERENCES

[1] C.C. Aggarwal and C. Zhai, *"Mining Text Data"*, MA: Springer, Eds. Boston, pp. 163-222, 2012.

[2] T. Jo, *"Classifying News Articles Using Feature Similarity K Nearest Neighbor"*, In Proc. of the 6th Int. Conf. on Green and Human Information Technology, ICGHIT 2018, Chiang Mai, Thailand, pp. 73-78, 2018.

[3] S. Xu, Y. Li and Z. Wang, "*Bayesian Multinomial Naïve Bayes Classifier to Text Classification*", In Proc. of the 11th Int. Conf. on Multimedia and Ubiquitous Engineering, MUE 2017, Proc. of the 12th Int. Conf. on Future Information Technology, FutureTech 2017, Seoul, South Korea, pp. 347-352 2017.

[4] A. Hassaine, S. Mecheter and A. Jaoua, *"Text Categorization Using Hyper Rectangular Keyword Extraction : Application to News Articles Classification",* In Proc. of the 15th Int. Conf. on Relational and Algebraic methods in Computer Science, RAMiCS 2015, Braga, Portugal, pp. 312-325 , 2015.

[5] A. A. Hakim, A. Erwin, K. I. Eng, M. Galinium and W. Muliady, *"Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF – IDF) approach",* In Proc. of the 6th Int. Conf. on Information Technology and Electrical Engineering, ICITEE 2014, Yogyakarta, Indonesia, pp. 1-4, 2014.

[6] S. Qiu, M. Jiang, Z. Zhang, Y. Lu and Z. Pei, *"Chinese News Text Classification of the Stacked Denoising Auto Encoder Based on Adaptive Learning Rate and Additional Momentum Item"*, In Proc. of the 15th Int. Symposium on Neural Networks, ISNN 2018, Minsk, Belarus, pp. 578-584, 2018.

[7] C. Li, G. Zhan and Z. Li, *" News Text Classification Based on Improved Bi-LSTM-CNN,"* In Proc. of the 9th Int. Conf. on Information Technology in Medicine and Education, ITME 2018, Hangzhou, China, pp. 890-893, 2018.

[8] M. Fanjin, H. Ling and T.J.W. Xinzheng, *"The Research of Semantic Kernel in SVM for Chinese Text Classifiction"* , In Proc. of the 2nd Int. Conf. on Intelligent Information Processing, IIP'17 2017, Bangkok, Thailand, 2017.

[9] A.D. Patel and Y.K. Sharma, *"Web Page Classification on News Feeds Using Hybrid Technique for Extraction"*, In Proc. of the 3rd Int. Conf. on Information and Communication Technology for Intelligent Systems, ICTIS 2018, Ahmedabad, India, pp. 399-305, 2018.

[10] F. Kabir, S. Siddique, M. R. A. Kotwal and M. N. Huda, *"Bangla text document categorization using Stochastic Gradient Descent (SGD) classifier",* In the Int. Conf. on Cognitive Computing and Information Processing, CCIP 2015, Noida, India, pp. 1-4, 2015.

[11] S. Brindha, K. Prabha and S. Sukumaran, *"A survey on classification techniques for text mining"* , In Proc. of the 3rd Int. Conf. on Advanced Computing and Communication Systems, ICACCS 2016, Coimbatore, India, pp. 1-5, 2016.

[12] M. Lichman, "*News Aggregator Data Set*", UCI Machine Learning Repository, 2013. (https://archive.ics.uci.edu)

Authors Profile

Mr. Feroz Ahmed is currently pursuing M.Tech in Computer Science & Engineering from Jamia Hamdard, New Delhi. He has completed his B.Tech in Information Technology from Jamia Hamdard only. He has presented a paper on Data Science in ICIDSSD' 19 .

Ms Shabina Ghafir pursued Bachelor of Technology from G.B.Pant University of Agriculture & Technology, Pantnagar (Nainital), India and Master of Technology from A.M.U. Aligarh (Uttar Pradseh), India in year 2006. She is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Science and Engineering,Jamia Hamdard, New Delhi since 2008. She is a life member of the ISTE since 2011. Her main research work focuses on Load balancing Algorithms,Cloud Computing, Software Engineering, Data Science, and Computational Intelligence based education. She has 12 years of teaching experience and 6 years of Research Experience.

