iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** II **Month of publication:** February 2020

DOI: http://doi.org/10.22214/ijraset.2020.2015

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 8 Issue II Feb 2020- Available at www.ijraset.com

Analysis on News Articles Classification using Machine Learning

Shreeka[1], Dr. Ravikumar G K[2], Mr. Prasanna Kumar M J[3]
[1]*PG student, Dept Of CSE, B.G.S Institute of Technology, B.G Nagar, Karnataka, India*
[2]*Professor and Head, Dept of R & D, B.G.S Institute of Technology, B.G Nagar, Karnataka, India*
[3]*Assistant Professor, Dept of CSE, B.G.S Institute of Technology, B.G Nagar, Karnataka, India*

Abstract: ***News articles order is an administered learning approach in which news articles are relegated classification names dependent on probability showed by a preparation set of marked articles. A framework for programmed classification of news articles into a standard arrangement of classifications has been executed. The proposed work will utilize Term Frequency–Inverse Document Frequency (TF-IDF) term weighting plan for improvement of grouping systems to get more advanced outcomes and utilize two administered learning draws near, i.e., Support Vector Machine (SVM) and K-Nearest neighbor (kNN) and think about the exhibitions of the two classifiers. Every news report is preprocessed and changed into a term-archive lattice. Besides, various methodologies, for example, stemming, stop word expulsion, include decrease have been executed for both execution and precision upgrades.***

Keywords: Machine Learning, NLP, SVM

I. INTRODUCTION

There is a tremendous measure of data that is developing at an exponential rate ordinarily as news stories. Digitization has prompted an expanding number of individuals exchanging to online hotspots for day by day news sources. Equivalent to the propels in the advanced time, individuals for the most part are engrossed with rushed work-life and like to peruse articles relating to their inclinations.

Robotized news stories arrangement or characterization is the programmed order of news stories or archives under predefined classifications or classes. It is one of the utilizations of content order. It is an administered learning approach. Content characterization is a sort of technique identified with Natural Language Processing (NLP). It finds social mode (classifier) between content's characteristics (highlight) and content's classification as indicated by a marked preparing content corpus, and afterward uses the classifier to order new content corpus. Content grouping can be partitioned into two sections: preparing and ordering. The motivation behind preparing is to structure classifier, which can be utilized to group new messages by the association between preparing content and classification. Ordering intends to make the obscure new content relegated with the realized class mark. Content characterization comprises of highlight determination, document representation, or highlight change, application and assessment of calculations applied. Content order can give calculated perspectives on report assortments and has significant application in the genuine word. AI strategies, for example, Naïve Bayes, Support Vector Machine, k-Nearest Neighbor, Decision Tree and others are applied to computerize grouping process. In managed text classification, the calculation is prepared on the marked information and its presentation results are estimated on the already inconspicuous test information. For preparing and furthermore testing the framework, the corpus is shaped from Azerbaijani news stories. Right now there are in excess of 150000 named articles in the corpus. Albeit a ton of research over content arrangement has been done, the exploration on Azerbaijani language is rare. That is the reason, we imagine that the examination that we have done will help specialists after us and will empower them to profit by the approach and the finishes of our exploration. This will likewise assist them with furthering gain ground in characteristic language preparing issues.

II. LITERATURE SURVEY

Trstenjak Bruno purposed a structure dependent on Term Frequency–Inverse Document Frequency (TF-IDF) approach for content order. The creators have given the likelihood of utilizing K-Nearest Neighbor (kNN) classifier with TF-IDF technique and purposed a system for content order, which empowers the grouping of different various classifications of source records as indicated by different parameters, estimation, and investigation of results. The structure was assessed based on the quality and speed of arrangement. The consequences of the analysis showed the positive and negative highlights of the calculation. Consequently, it has given the course to the advancement of comparable sort of frameworks [1].

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 8 Issue II Feb 2020- Available at www.ijraset.com

In other paper have used kNN-based learning approach for text categorization, in the first step, and the documents have been categorized using kNN classifier, and then compared with Naïve Bayes classifier and Term graph by returning the most relevant documents. In this paper, the authors concluded that kNN showed maximum accuracy as compared to the Naive Bayes and Term Graph [2].

Rahmawati Dyah has proposed an approach for multi-label classification for Indonesian news articles. In the research work, two approaches, i.e., problem transformation and algorithm adaptation were investigated. The paper focused on four factors, i.e., feature weighting, feature selection method, multi-label classification, approach and single-label classification approach. The experimental results showed that the Support Vector Machine (SVM) algorithm performed the best when it was used with TF-IDF [3].

III. WORKING PRINCIPLE

The steps of the research methodology are explained as follows:

Input training and testing new documents into the system.

Processing of the input news documents is done, i.e., removing stop words, whitespaces, punctuations, numbers, and performing stemming of training and testing documents etc.

Document Term Matrix of preprocessed news documents is created using the TF-IDF approach as purposed in this research.

Classification of testing news documents to different categories, i.e., Sports, Business, and Science and Technology is performed using the classifiers SVM and kNN. The performance of kNN can be improved by changing the value of k.

There are many different ways of measuring the performance of a classifier and evaluate its performance. For supervised classification with two classes, each and every performance measure is based on four numbers obtained after applying the classifier to the test dataset. These numbers are called true positives tp, false positives fp, true negatives tn, and false negatives fn.

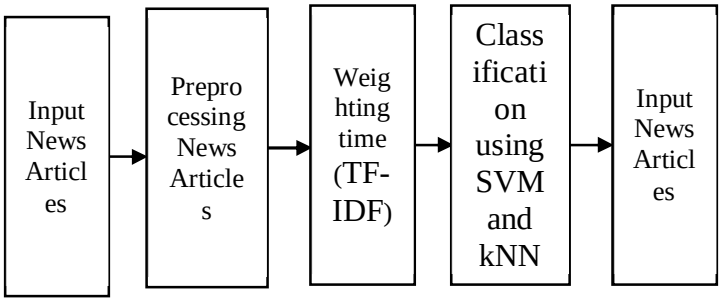


Fig.1. Block diagram

A. Data Retrieval

In this stage, news story information is being recovered from different news sites. Fig.2 shows the procedure of information recovery. The initial step includes Parsing of RSS channels from news sites. During RSS channel Parsing, connections of different articles are being separated. The second step of information recovery process is gathering the recovered URLS in a record which is to be utilized for further preparing. The third step includes bringing of article content information from the gathered URLS. Every URL is visited and article content is separated from the HTML page of every news article. During this procedure RSS channels for the 5 urban areas Delhi, Chandigarh, Kolkata, Mumbai and Lucknow were crept from three news site Indian Express, Hindustan Times and Times of India. A sum of 2000 articles was gathered with a conveyance with the end goal that the informational collection has 400 articles for each city. The equivalent conveyance of dataset is guaranteed so that there is no predisposition in the preparation stage.

B. Text Pre-processing

This stage as appeared in Fig.2 includes pre-handling of the got information. The initial step includes tokenizing the articles in which the succession of characters is changed over into grouping of string which have distinguished significance. When the articles are tokenized the following stage is stemming where each word in the article is diminished from its inflectional or derivationally related structure to its base form. It is trailed by expulsion of stop words as these are the most widely recognized words and are of little noteworthiness in this arrangement procedure. For the expulsion of stop words, a rundown of explicit stop words was made to be expelled separated from the stop words gave by the nltk python bundle. The last advance of this procedure is the Part of Speech Labelling where each word in the article is relegated a piece of discourse, for example, thing, and action word, and modifier, thing plural and so forth.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 8 Issue II Feb 2020- Available at www.ijraset.com

C. Training a Classifier

Preparing a classifier initially includes the pre-handling module which separates significant piece of the article. This progression is basic to improve the exactness of the classifier. Contribution to the classifier is the preparation set and the arrangement of names comparing to it. The handled news stories are named numerically dependent on the city label which is pre-chosen. Since contribution to the classifier is two vectors, the arrangement of news stories what's more, the marks should be vectorized. After vectorization of these two elements, they go about as a contribution to the classifier. As it were 80% of the dataset is utilized for preparing the classifier; the rest is used for testing. Since preparing is performed just once, the classifier object which is prepared is put away in a sticky situation document. Pickling is done to serialize the article and putting away it into the circle for the testing stage. Comparable procedure is applied for dumping the Count Vectorizer object for additional utilization.

D. Testing a Classifier

The pickle document containing the put away classifier and the Check Vectorizer object is stacked. Since the put away classifier is prepared, trying information is nourished into it. The classifier predicts the class, for this situation city, of the relating article. The grouping as performed by the prepared classifier, acts a premise for deciding the exactness of the model. In the wake of testing, the exactness of the classifier is noted dependent on different execution measurements like Precision, Recall and F1 score.

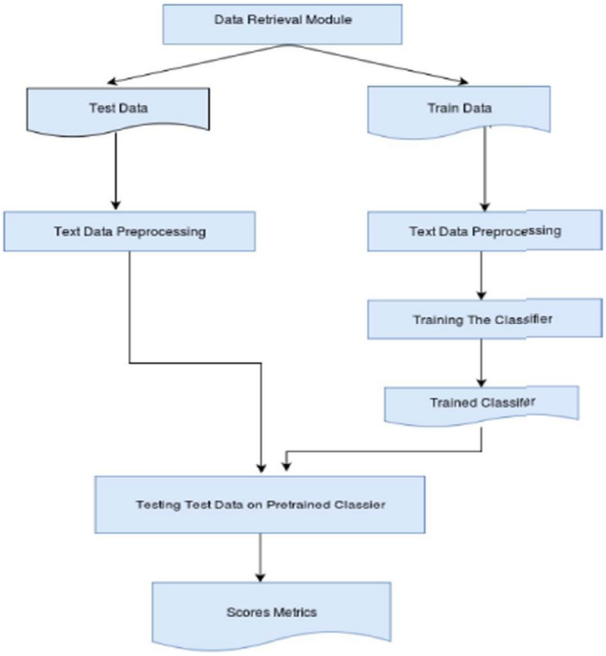


Fig.2. Flow chart of the process

E. Naive Bayes

Naive Bayes classifier is a popular method for text classification problems where given a document or article the
Classifier has to decide the category of the article. It was first proposed by D.Lewis [4]. Naive Bayes classifier is based on probabilistic technique of classification which derives its roots from Bayes Theorem. It is based on the assumption of independence between the various features. It is a scalable classifier and can run efficiently with large data sets. Naive
Bayes classifier is fast as compared to other classifiers and thus is used as a baseline for text classification problems.

F. SVM Classifier

SVM (Support Vector Machine) [5] works on the principle of Supervised Learning. SVM requires a training set and labels associated with it. After training, if a test data is fed in, the model assigns it to one category or the other. It performs well with linear classification. It can even work efficiently on a nonlinear classification using a kernel trick by mapping the inputs into high dimensional feature space. It constructs a hyperplane for classification. The hyper-plane is chosen such that the distance between the nearest data point on either side is maximized.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 8 Issue II Feb 2020- Available at www.ijraset.com

IV. PROS OF NEWS CLASSIFICATION

News article classification is digitalized using machine learning provides better accuracy. User can easy classify the article based on their interest.

V. CONCLUSION

In this paper, we have researched the likelihood to utilize AI calculations to characterize the news articles based on algorithms. The investigations show that this issue can be effectively comprehended by utilizing different Classifiers, for example, Naive Bayes, Support vector Machines and KNN. Arbitrary Timberland has outflanked different classifiers. Bayes has performed well as well and Support Vector machine is at the base as far as the exhibition measurements utilized in our methodology. The proposed framework can be utilized as a piece of something else complex news article order frameworks.

VI. ACKNOWLEDGMENT

I would like to express my sincere gratitude towards my guide Dr. Ravikumar G K, Professor and head, Dept of R & D, BGSIT, for the help, guidance and advice in development of this methodology. I would like to express my sincere gratitude towards Mrs. Divya B M, Asst. Professor, BGSIT and, Professor, BGSIT, for the support and guidance.

REFERENCES

[1] Trstenjak, B., Mikac, S., Donko, D.: kNN with TF-IDF based framework for text categorization. Proc. Eng. 69, 1356–1364 (2014). In: 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013

[2] Bijalwan, V., Kumar, V., Kumari, P., Pascual, J.: kNN based machine learning approach for text and document mining. Int. J. Database Theory Appl. 7(1), 61–70 (2014)

[3] Rahmawati, D., Khodra, L.M.: Automatic multilabel classification for Indonesian news articles (2015). IEEE 978-1-4673-8143-7/15

[4] Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In: Data Mining and Knowledge Discovery Handbook, pp. 667–685. Springer, Berlin (2010)

[5] Ikonomakis, M., Kotsiantis, S., Tampakas, V.: Text classification using machine learning techniques. WSEAS Trans. Comput. 4(8), 966–974 (Aug 2005)

[6] B. Pendharkar, P. Ambekar,P. Godbole, S. Joshi, and S. Abhyankar,"Topic categorization of rss news feeds," Group vol. 4, p. 1, 2007.

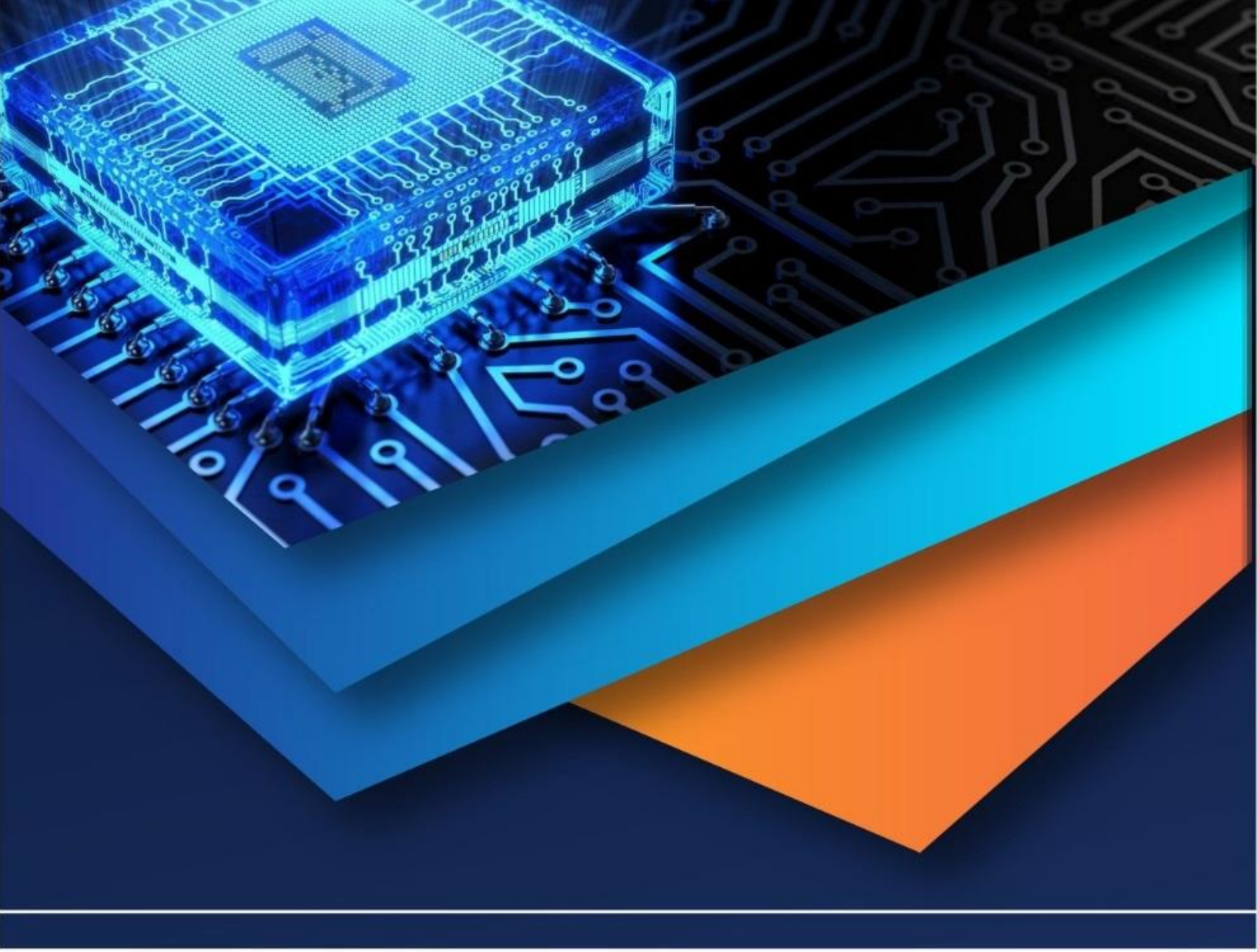
[7] D. Shen, Z. Chen, Q. Yang, H. Zeng, B. Zhang, Y. Lu, and W.Ma,Web-page classification through summarization, in Proceedings ofthe 27th annual international ACM SIGIR conference on Research anddevelopment in information retrieval. ACM, 2004, pp. 242249

[8] Leo Breiman, Random forests, Machine Learning. vol. 45, no. 1, pp.532, 2001.

[9] D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," Machine Learning: ECML-98, pp. 415, 1998

[10] J,.K. M. Han, Data Mining:Concepts and Techniques, 2nd ed. 2006.

[11] H. a. K. S. Yu, "SVM tutorial: Classification, regression, and ranking," Handbook of Natural Computing 2009.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY