International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 2454–2460

Multi-Label Text Classification of News Article

V.Srividhya[1], P.Megala[2]

[1]*Department of Computer Science,*

Avinashilingam Institute for Home Science and Higher Education for Women Coimbatore, India

[2]*Department of Computer Science,*

Avinashilingam Institute for Home Science and Higher Education for Women Coimbatore, India

[1]*vidhyavasu@gmail.com*
[2]*p.megala05@gmail.com*

Abstract

The paper entitled "Multi-Label Text Classification of News Articles". The significant goal of current paper is classify the text of News articles using machine learning algorithms - Logistic Regression, Random Forest, XG Boost and Naive Bayes algorithms and compare them to discover the most suitable approach. Multi-Label Text Classification is a classification task which consists of more than two groups; each label are mutually exclusive. The use of Text classification models is to classify text into sort out as groups. Text classification is also known as text tagging or text categorization. In general, an automatic document classification algorithm provides a predetermined label to the text documents (test dataset) on the basis of classifier model progressed using the supervised machine learning algorithm. This work focuses on Logistic Regression, Random Forest, XG Boost and Naive Bayes algorithms which are part of supervised machine learning algorithms in automatic classification of text documents. This work is consists with the five stages. The initial part is loading the data sets and explores them. The second phase is Pre- processing. The third phase is Vector Space modeling which is used to represent documents in the format as vectors of identifiers and model fitting is fourth part with four machine learning algorithms. The fifth phase is Performance measure that describes the accuracy algorithms to find which algorithms works best. To carry out this work, BBC news article dataset is collected form Insight Resources.

Keywords: *Classification, articles, vector space model, machine learning algorithm.*

1. Introduction

Massive volumes of text are available format to users online such as scientific articles, news, product reviews and social media content, etc. It is more practical for users to look for information by browsing through categories rather than searching the whole information space. True uses of text arrangement frequently require a framework to manage a huge number of classifications characterized over an enormous scientific categorization. Since building these content classifiers by hand is tedious and exorbitant, mechanized content order has picked up significance throughout the years. Classification methods developed based on machine learning deal with text classification problem in a fantastic way.

Automatic text categorization is gaining significance of text with the growth in electronic format grows. It is imperative to have efficiency in accessing these documents.

This task is considered challenging because grouping textual data effectively as for the context in which it is used is not much clear as compared to numerical data. Examples of textual data

ISSN: 2233-7857 IJFGCN

applications that can benefit through categorization involves technical and journal articles, online newspapers, books, manuals, electronic mail, memos, and so on.

Text classification is crucial in the domain named text mining and is classified as Topic-Based Text Mining and Genre-Based Text Mining. The Documents are classified on the basis of topics in text-based categorization. The text database has been growing quickly due to the rapid development amount of information that is available as electronic publications, from e-mail and World Wide Web. Currently, most information from the government, business, industries and institutions are being stored electronically as text databases. The text database contains mostly semi- structured data as they are not unstructured or completely structured. Multi-Label Text Classification (MLTC) contains data with an enormous amount of features and hence results in the risk of great dimensionality. Furthermore, the presence of irrelevant and redundant features complicates the MLTC by generating ambiguous data representation and poorly describing category labels. Feature Extraction (FE) is commonly applied to unfold this challenge by reducing the original large feature space and retaining the relevant features. Moreover, the accuracy of MLTC is largely driven by effectiveness of FE for representing various features.

2. Related Work

The text mining contemplates are gaining more significance as of late due to the accessibility of the expanding number of the expansion of electronic records from an assortment of sources which include unstructured and semi structured information. The techniques of machine learning cooperate to naturally order and find designs from the various kinds of the archives [1].

The huge amount of comparative studies are also done in document categorization. Such as the author used KNN, NB and Term Gram for this task. They showed a comparative study where the efficiency of KNN is a better choice and Term Gram [10].

In this paper a relative learning on DT, KNN, Rocchios Algorithm, Back propagation, NB and SVM has been done. Here for 20 new group's dataset they showed SVM performed far better than all the other approaches they have used [4].

News article suffers from a lot of ambiguity during classification due to its various matching categories and the weak reliability indices offered by some classification systems often employed [2].

TF-IDF increases when the term exists in small number of documents that distinguish a particular document. At the same time it increases when the term appears many times in the document [3].

Paper that authors used Reuter's group as dataset to classify documents to ten classes. In this paper they make comparison between different degree of SVM polynomial and SVM with RBF function. The outcomes picked up for polynomial SVM was 80% and for SVM with RBF 80.4% they likewise contrasted and another model however for the most part paper is about SVM [8].

Naïve bias method is kind of module classifier under known priori probability and class conditional probability. It is fundamental idea is to compute the possibility that text D is fit in to group C. The multivariate Bernoulli and multinomial model are two models are available for naive Bias. Out of this model multinomial model is progressively reasonable when database is huge [6].

In a different setting, multi-class classification is tried by combining kernel density estimation with k-NN. It improves the weighting rule of k-NN, thereby increasing the precision of classification. It has also been proven efficient for complex classification problems [9].

In this paper authors developed a classification model for categorization of cricket sports news. They used SVM which was based on Lib SVM and got best performance [7].

ISSN: 2233-7857 IJFGCN

Statistical topic modeling is applied for multi- label document classification, where each document gets assigned to one or more classes. It became an interesting topic in the past decade as it performed well for datasets with increasing number of instances for an entity [5].

3. Methodology

The methodology is constructed in to five phases. Each phase has its own task which is followed by the other. First phase is the Load and Explore the dataset phase, second phase is the preprocessing phase, third phase is Vector Space modeling which represents text as vectors of identifiers and applying Text classification Algorithms is the fourth phase and Performance measure of algorithms is the fifth phase. The following figure 1 shows the overall methodology diagram.

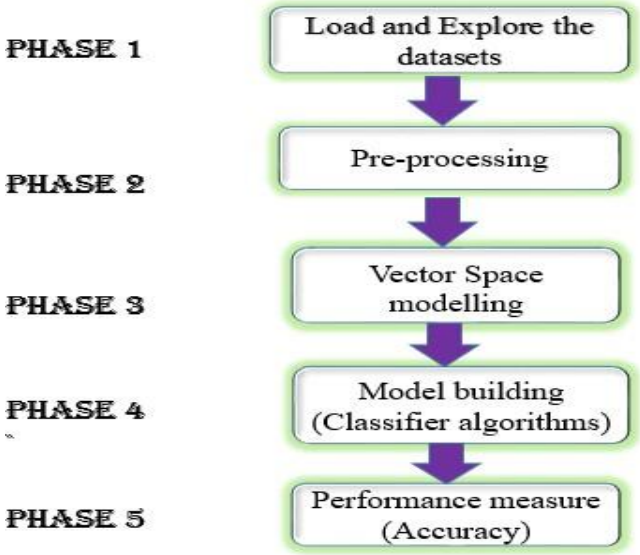

Figure 1. Overall Methodology Diagram

A. LOAD AND EXPLORE THE DATASET - BBC NEWS ARTICLES DATASET (2225 Records)

British Broadcasting Corporation (BBC) news data set (Greene & Cunningham, 2006) is a raw text documents which contains 2225 text files from its website relating to stories in five effective areas from 2004-2005. The text documents are arranged into five folders with the following labels such as entertainment, business, sports, politics, tech and each of them contains new articles related to that class label. Presentation of information and data graphically is known as Data visualization. Categories of news articles are explored using bar chart to visualize the number of articles that exists in the dataset.

B. PRE-PROCESSING

The first and foremost step in the processing of text is preprocessing. The significance of pre-processing is utilized in almost every created frame work related with text handling and natural language processing. This phase includes words identification, sentences identification, stop words elimination and stemming. Stemming is the process used to eliminate words which are reproduced from original ones. Preprocessing phase is used to reduce the size of text. The steps performed in text classification are tokenization, stop words removal, Removal of punctuations, Removal of integers and numbers, Removal of extra spaces and Stemming.

C. VECTOR SPACE MODELLING

Representation of text documents as vectors of identifiers can be done through vector space model which like index terms. In this each individual token number considered to be an aspect.

ISSN: 2233-7857 IJFGCN

TF-IDF (term frequency-inverse document frequency) is a factual measure that assesses how significant a word is to a document in a accumulation of documents. When the term comes more in the document it gains more importance logically. Vector can used to represent the document in bag of words model.

Hence in this, two matrices have to be computed, one containing the inverse document frequency of a word in the entire corpus of text collection and another containing the term occurrence of each word in each document.

Text arrangement is the assignment of consequently arranging a lot of reports into classifications from a predefined set.

D. MODEL BUILDING USING MACHINE LEARNING

Text categorization is the assignment of automatically arranging a group of text into classifications from a predefined set. This research area combines of information retrieval (IR) technology, Machine Learning (ML) and Data mining technology together. Generally machine learning consists of two key aspects, modeling and optimization. Modeling refers to the way of designing the distinction between two limits or distribution of the provided training set can be made with the probability. It should be noted that the effective methods are used to identify the most suitable metrics of the selected model. Machine learning is relevant to domains such as artificial neural networks, pattern recognition, information retrieval, artificial intelligence, data mining, and function approximation.

i. LOGISTIC REGRESSION

The commonly used classification method is Logistic Regression, which is suitable to multi class problem that contains two or more possible results. For a given set of independent variables, dependent variables are categorically distributed using Logistic Regression model.

ii. RANDOM FOREST CLASSIFIER

The random forest method combines individual decision trees into large ensembles. Each tree has the same weight in voting (fair voting scheme). Random forests can deal with high dimensions and distinguish feature relevance which makes them a suitable tool for text categorization.

iii. XGBOOST

An algorithm called XGBoost is used commonly in the arena of Kaggle competitions and machine learning mainly for structured data. It competes with the gradient boosted decision trees which are used for the operations faster and better manner.

iv. NAÏVE BAYES TEXT CLASSIFIER

It is a simple probabilistic classifier based on applying Bayes' Theorem with strong independence assumptions. It treats text as bag of words and the occurrence of terms and their positions are independent of the probabilities.

E. PERFORMANCE MEASURE

This is final step of Text categorization, in which text classifiers are evaluated experimentally, instead of analytical evaluation. To evaluate each model and high accuracies obtained as results for full text articles, in this work the k-fold cross-validation technique is used iteratively training the model on distinct subcategories of the data and testing against the held-out data.

4. Results and Discussion

In this work, BBC news articles dataset is used. The data set is downloaded from UCI Benchmark repository. British Broadcasting Corporation (BBC) news data set (Greene &Cunningham, 2006) is in the structure of raw text documents and contains 2225 files from the British Broadcasting Corporation website. The text documents are arranged into five folders named with the class label (business, entertainment, politics, sport, and tech) and each of them

ISSN: 2233-7857 IJFGCN

contains new articles related to that class label is shown in figure 2. The figure 3 shows the number of records in BBC news dataset based on different labels using bar plot.

A1 | category

	A	B
1	category	text
2	tech	tv future in the hands of viewers with home theatre systems plasma high-definition tvs and digital video recorders moving int
3	business	worldcom boss left books alone former worldcom boss bernie ebbers who is accused of overseeing an $11bn (Â£5.8bn) fraud
4	sport	tigers wary of farrell gamble leicester say they will not be rushed into making a bid for andy farrell should the great britain rugl
5	sport	yeading face newcastle in fa cup premiership side newcastle united face a trip to ryman premier league leaders yeading in the f
6	entertainment	ocean s twelve raids box office ocean s twelve the crime caper sequel starring george clooney brad pitt and julia roberts has g
7	politics	howard hits back at mongrel jibe michael howard has said a claim by peter hain that the tory leader is acting like an attack mong
8	politics	blair prepares to name poll date tony blair is likely to name 5 may as election day when parliament returns from its easter break
9	sport	henman hopes ended in dubai third seed tim henman slumped to a straight sets defeat in his rain-interrupted dubai open quart
10	sport	wilkinson fit to face edinburgh england captain jonny wilkinson will make his long-awaited return from injury against edinburgh
11	entertainment	last star wars not for children the sixth and final star wars movie may not be suitable for young children film-maker george luca
12	entertainment	berlin cheers for anti-nazi film a german movie about an anti-nazi resistance heroine has drawn loud applause at berlin film fest
13	business	virgin blue shares plummet 20% shares in australian budget airline virgin blue plunged 20% after it warned of a steep fall in full
14	business	crude oil prices back above $50 cold weather across parts of the united states and much of europe has pushed us crude oil prices
15	politics	hague given up his pm ambition former conservative leader william hague says he will not stand for the leadership again havi
16	sport	moya emotional after davis cup win carlos moya described spain s davis cup victory as the highlight of his career after he beat ar
17	business	s korean credit card firm rescued south korea s largest credit card firm has averted liquidation following a one trillion won ($960

Figure 2. BBC News Article Dataset

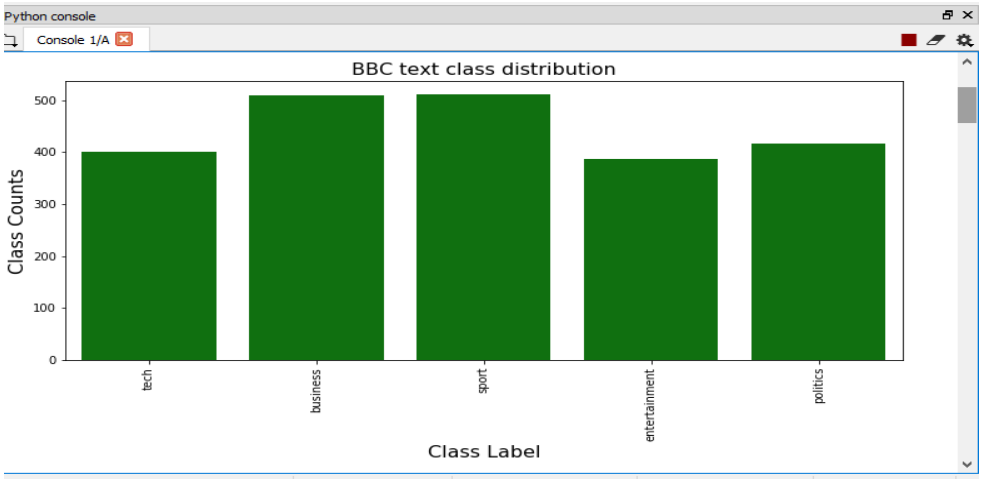

Figure 3. Visualizing the BBC News Article Dataset

Accuracy as a metrics used to estimate the performance of classification models. Everyone use classification accuracy in order to the gauge the level of performance of the constructed model. Accuracy refers to number of predictions that are correct out of all predictions. The approach used to conduct the experiments on the data set using numerous machine learning algorithms they are Logistic Regression, Random Forest, XGBoost and Naive Bayes algorithms. Out of many machine learning algorithms, four were used on BBC news data set to see if classification task works well. The accuracy result of various text classifier for BBC News dataset is shown in the below table - Table 1.

ISSN: 2233-7857 IJFGCN

Table 1: Performance metrics of classification algorithms - BBC News Articles

Performance Metrics	
Classifier Model	**Accuracy**
Logistic Regression	89%
Random Forest	85%
XGBoost	93%
Naive Bayes	97%

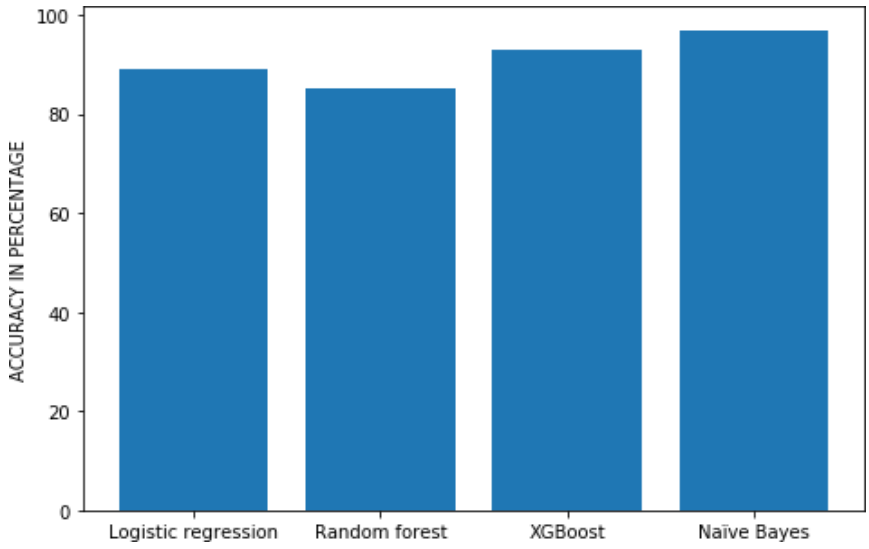

Figure 4: Accuracy of algorithms based on BBC News Articles

The above Table 1 shows the accuracy value of Logistic Regression, Random Forest, XGBoost and Naive Bayes classifier which is applied on BBC News article dataset. Among these four algorithms Naive Bayes classifier has highest accuracy when compare to other algorithms shown in figure 4.

5. Conclusion

Due to the umpteen number of textual information available on web which can't be analyzed by humans, a service oriented approach can be useful for retrieving important information from text documents. Logistic Regression, Random Forest, XGBoost and Naive Bayes algorithms, which are parts of supervised machine algorithm widely used in automatic classification of News articles. From several classification experiments conducted on BBC news article dataset become evident that with sufficient amount of data and decent number of features (words), high accuracies can be achieved for text classification tasks. Different machine learning classification algorithms are applied in BBC News to find the better accuracy. From those algorithms, Naïve Bayes classification algorithm gives the highest accuracy. The experiments revealed that the most reliable categorization can be reached with the Naïve Bayes classifier.

This work can be extended by using unsupervised learning algorithm for text classification and comparing supervised text classification algorithms with semi supervised and unsupervised algorithm. Also, this project can be continued for other languages to make classifier more versatile.

ISSN: 2233-7857 IJFGCN

References

[1] F. Sebastiani, “Text categorization”, Alessandro Zanasi (ed.) Text Mining and its Applications, WIT Press, Southampton, UK, pp. 109- 129,2005.

[2] Gurmeet Kaur and Karan Bajaj (2016). News Classification and Its Techniques: A Review. IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 1,, PP22-26.

[3] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, And Lofti Zadeh, "Feature Extraction" Usa 2006.

[4] P. Y. Pawar and S. Gawande, “A comparative study on different types of approaches to text categorization, ”International Journal of Machine Learning and Computing, vol. 2, no. 4, p. 423,2012.

[5] Rubin, T. N., Chambers, A., Smyth, P., &Steyvers, M.. “Statistical topic models for multi-label document classification. Machine Learning”, 88,(2012), pp. 157–208.

[6] SHI Yong-feng, ZHAO, “Comparison of text categorization algorithm”, Wuhan university Journal of natural sciences, 2004.

[7] T.Zakzouk and H.Mathkour, “Text classifiers for cricket sports news”, in Proceedings of International Conference on Computer Communication and Management (ICCCM 2011).

[8] Thorsten Joachims, "Text Categorization With Support Vector Machines: Learning With Many Relevant Features" Germany2003.

[9] Tang, X., &Xu, A. (2016). “Multi-class classification using kernel density estimation on K-nearestneighbours”, Electronics Letters, 52, 8, 600–602.

[10] V. Bijalwan, V. Kumar, P. Kumari, and J. Pascual, “Knn based machine learning approach for text and document mining,” International Journal of Database Theory and Application, vol. 7, no. 1(2014), pp. 61–70.

[11] S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy in Magnetism” ,Vol.III, G.T.Rado and H.Suhl,Eds. New York: Academic, (1963), pp.271–350.

ISSN: 2233-7857 IJFGCN