Book Genre Classification using ML Algorithms

Vraj Patel[1], Meet Soni[2], Janvi Patel[3]
Department of Information and Communication Technology
Parul Institute of Technology (PIT), Vadodara, Gujarat, India
Email: *170305114005@paruluniversity.ac.in, 170305114006@paruluniversity.ac.in*
DOI:- https://doi.org/10.47531/SC.2022.22

Abstract

Rapid progress in digital data acquisition techniques have led to huge volume of data. More than 80 percent of today's data is composed of unstructured or semi-structured data. The recovery of similar patterns and trends to see the text data from huge volume of data is a big issue. Text mining is a process of extracting interesting and nontrivial patterns from huge amount of text documents. There lie many techniques and tools to mine the text documents and discover the information for future and process in decision making. The choice of selecting the right and appropriate text mining technique helps to recover the speed and slows the time and effort required to get valuable information. This paper briefly discusses and analyzes the text mining techniques and their applications. With the advancement of technology, more and more data is available in digital form. Among which, most of the data (approx. 85%) is in unstructured textual form. Thus, it has become essential to build better techniques and algorithms to get useful and interesting data from the large amount of textual data. Hence, the field of information extraction and text mining became popular areas of research, to get interesting and needful information

Keywords: - *Text Classification, Book Classification, K-Nearest Neighbor, Support Vector Machine, Naïve-Bayes, Convolutional Neural Network*

INTRODUCTION

In this world there are many people who like to read books and there are lots of books in the world so sometimes people can't identify books by its cover that book is as per his/her requirement or not to solve this major problem we building one model which classify the book genre by its name so that people can easily identify book that is fulfill his/her requirement or not. To build this model we are going to use some text classification techniques and some ML algorithm. The text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. Which include unstructured and semi structured information. The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification (supervised, unsupervised and semi

supervised) and summarization Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify the documents and discover patterns from different types of the documents . Text classification (TC) is an important part of text mining, looked to be that of manually building automatic TC systems by means of knowledge-engineering techniques, i.e. manually defining a set of logical rules also called as training , that convert expert knowledge on how to classify documents under the given set of categories. For example would be to automatically label each incoming news with a topic like “sports”, “politics”, or “business”. A data mining classification task starts with a training set D = (d1….. dn) of documents that are already labeled with a class C1, C2 (e.g. sport, politics). The task is then to determine a classification model which is able to assign the correct class to a new document d of the domain. Basically there are two stages involved in Text Classification.

Training stage and testing stage. As explained in the above paragraph, in training stage documents are preprocessed and are trained by a learning algorithm to generate the classifier. In testing stage, a validation of classifier is performed. There are many traditional learning algorithms to train the data, such as Decision trees, Naïve-Bayes (NB), Support Vector Machines (SVM), kNearest Neighbor (kNN), Neural Network (NNet), etc.

In this research, we study the problem of text classification, that is classifying the news documents into different categories based on three different supervised algorithms namely Naive Bayes classifier, Vector Space Model for text classification and a new technique -Use of Stanford Tagger for text classification. We have tried to compare the efficiency and accuracy of the algorithms to analyze the effectiveness of each algorithm. The research has been carried out on two different datasets namely 20 Newsgroup and New Dataset of news for five categories.

OBJECTIVES AND SCOPE OF THE STUDY

Our final moto is to create a model that can determine how representative a title is of its genre and by the way, it is very difficult for even a human to distinguish between books of different categories. So, we are going to create a model that uses some Machine Learning Algorithm and text classification techniques which classify the book Genres by its Title.

Our system will be useful for Libraries, Book stores, Schools, Colleges, Digital Libraries etc. With the dramatic increase in the amount of content available in digital forms gives rise to a problem to manage this data. As a result, this research can be useful for classification and manage the data. The system will improve as per feedback received from users.

RESEARCH METHODOLOGY

Many research papers are studied by us. Then we found many classification algorithms and feature selection methods and their advantages, disadvantages and limitations which we can use in our model to predict. We also found the required list of the algorithms are Naïve Bayes (NB), Support Vector Machine (SVM), K- Nearest Neighbor (K-NN), Convolutional Neural Network (CNN) and the feature selections are Information Gain (IG), Principal Component Analysis (PCA), Chi2 square (X (2)).

WORKING

First of users will register in our system. In the system there will be modules of old users and new users. If the user is new his/her details will be filled in the form and store in our database. If the user is old his/her details will be retrieved from database. The users will able to select genre from drop down menu and custom input box will be provided for the selection of genre which are not in the list. For different genres the list of various books will be shown.

We require to create a database in which we need to insert collections of books details and then in the model there is different machine learning algorithms which will be used as classifiers. Then the process starts where the details of new book will be entered as a raw input and the model will process that raw data entered and as a result it will show the book categorized by genre.

MACHINE LEARNING ALGORITHMS

1) K-Nearest Neighbor (K-NN) KNN is a classification algorithm is used for text classification. As given in (KNN classifies dataset or objects by voting several labeled training data with their smallest distance from each dataset or object. It uses the local neighborhood to predict the class of an object. The majority vote of its neighbors decides the class of an object. The object is assigned to the class most common among its k nearest neighbors. Here k is a positive integer. If k= 1, the object is assigned to class of that single nearest neighbor. The classes of these neighbors are decided using the similarity of each neighbor to new document vector, where similarity may be measured by for example the Euclidean distance or the cosine between the two document vectors.

Advantages

- The cost of the learning process is zero
- No necessity of assumptions about the characteristics of the concepts to learn have to be done
- It is very simple.

Disadvantages

- The model cannot be interpreted.
- It is computationally expensive, requires more time to find the k nearest neighbors when there is large number of training datasets.
- It has to compute distance of each test objects with whole training dataset.

2) Naïve-Bayes

Naïve Bayesian is simple and efficient to implement as it assumes that all the words of the documents are independent to one another. The Naïve Bayes Classifier is the simplest probabilistic classifier used to classify the text documents. Naïve Bayes method is kind of module classifier under known priori probability and class conditional probability. The basic idea is to use the joint probabilities of words and categories to estimate the class of a given document. Given a document di , the probability with each class cj is calculated as P(cj/di)=P(di/cj).P(cj)/P(di) As P(di) is the same for all class, then label(di) is the class (or label) of di , can be determined by label(di)= argMaxcj{ P(cj /di)} = arg Max{P(cj))/ P(di /cj)P(cj)} This technique Classify using probabilities and assuming independence among terms P(C/Xi XjXk) = P(C) P(Xi/C) P(Xj/C) P(Xk/C).

Algorithm A) Training Phase **Step 1:** Training System a) Applying Preprocessing methods for the data present in each categories. i.e. stop word

removal, Stemming. b) Tokenizing the data and storing the words along with its category in the database.

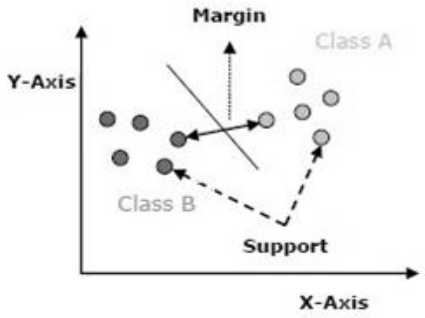
Step 2: Probability Calculations a) For each unique word in the categories, we try to find out the probability of each unique words belonging to that particular class. b) Formula for the probability is as follows: PrS[i] = Probability that word belongs to Document/class A (any category). PrC[i] = Probability that word belongs to Document/class B (any category). PrS[i] = [freq[i]/ freq[i]+freq2[j]]. PrC[i] = [freq2[j]/ freq[i]+freq2[j]]. c) Calculate probabilities for each category and store it in database. B) Testing Phase Step 1: Applying Pre-processing methods for the data present in test document. I.e. stop word removal, stemming.

Step 2: Tokenizing the data and storing the words along with its category in real time memory.

Step 3: Checking each unique word from test document with the word probability stored in database. If that word occurs in that particular category then probability of that word is added and this is repeated for all the words in that test document. Step 4: Probabilities of each category is calculated and the one with the highest probability is the correct match.

3) Support Vector Machine (SVM) Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. In 1960s, SVMs were first introduced but later they got refined in 1990. SVMs have their unique way of implementation as compared to other machine learning algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables.

An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).



The followings are important concepts in SVM

Support Vectors − Data points that are closest to the hyperplane is called support vectors. Separating line will be defined with the help of these data points.

Hyperplane − As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.

Margin − It may be defined as the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.

The main goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH) and it can be done in the following two steps −

- First, SVM will generate hyperplanes iteratively that segregates the classes in best way.
- Then, it will choose the hyperplane that separates the classes correctly.

4) Convolutional Neural Network (CNN)

Convolutional Neural networks are designed to process data through multiple layers of arrays. This type of neural networks is used in applications like image recognition or face recognition. The primary difference between CNN and any other ordinary neural network is that CNN takes input as a two-dimensional array and operates directly on the images rather than focusing on feature extraction which other neural networks focus on.

The dominant approach of CNN includes solutions for problems of recognition. Top companies like Google and Facebook have invested in research and development towards recognition projects to get activities done with greater speed.

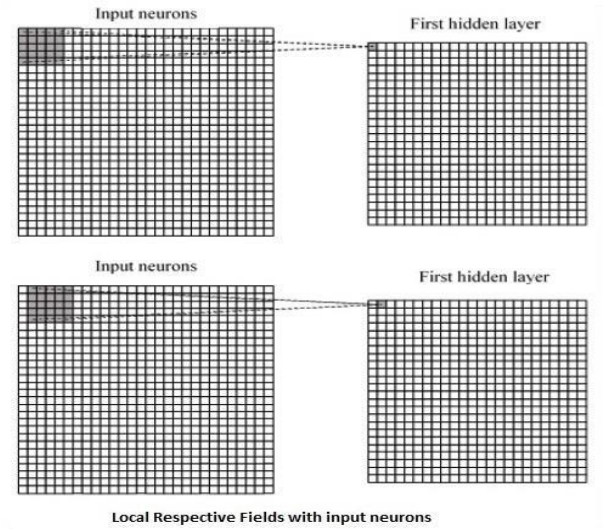
A convolutional neural network uses three basic ideas −

- Local respective fields
- Convolution
- Pooling

Let us understand these ideas in detail.

CNN utilizes spatial correlations that exist within the input data. Each concurrent layer of a neural network connects some input neurons. This specific region is called local receptive field. Local receptive field focuses on the hidden neurons. The hidden neurons process the input data inside the mentioned field not realizing the changes outside the specific boundary.

Following is a diagram representation of generating local respective fields −



Local Respective Fields with input neurons

If we observe the above representation, each connection learns a weight of the hidden neuron with an associated connection with movement from one layer to another. Here, individual neurons perform a shift from time to time. This process is called “convolution”.

The mapping of connections from the input layer to the hidden feature map is defined as “shared weights” and bias included is called “shared bias”.

ADVANTAGES OF BOOK CLASSIFICATION

Useful for libraries where all can manage books easily can be used in book stores where seller and consumer can find any book easily by its category. Also useful in E-Libraries. Using our system, the user gets classified books by genre to genre, this will save their time by not classifying manually.

DRAWBACKS

Naïve Bayes Classifier is limited by data scarcity for which any possible value in feature space, a likelihood value must be estimated by a frequentist. In Support Vector Machine (SVM) Lack of transparency in results caused by a high

number of dimensions (especially for text data). Computational of K-Nearest Neighbor model is very expensive and difficult to find optimal value of k. Finding a meaningful distance function is difficult for text data sets. Fails to capture polysemy and also still semantic and sentatics is not solved. The lack of transparency in the results. Naïve Bayes Classifier method makes a strong assumption about the shape of the data distribution. Continuous upgrading data & User Acceptance is also required which is an important prospect.

CONCLUSIONS

This paper has stated that classification of documents is one of the most fundamental problems in the machine learning and data mining. With the drastic increase in the world digitization, there has been an explosion in the volume of documents. Text Classification is hence needed to classify the documents according to the predefined classes based on their content. A comparative study has been done among different techniques which are used for classification such as nearest neighbor classifiers, SVM classifiers, neural networks, Naïve Bayes Theoram. When compared it was found that K-nearest neighbor algorithm (KNN) is the simplest method for deciding the class of the unlabeled documents and is a popular nonparametric method. But for the high dimensions, this method is not suitable for such documents. SVMs and Neural Network tend to perform much better when dealing with multi dimensions. For SVMs and Neural Network, large sample size is required to achieve maximum accuracy of the classifier, whereas Naïve Bayes may need a relatively less dataset and require little storage space. KNN, Neural Network is generally considered intolerant of noise; where association based classification and decision trees are considered resistant to noise because their pruning strategies avoid over-fitting the noisy data. Compared to other classifiers, SVM performs better as it has high accuracy, high speed of learning, high speed of classification, high tolerance to irrelevant features and noisy data than other classifiers. But still it seems difficult to recommend any one technique as superior to others as the choice of a modeling technique depends on organizational requirements and the data on hand.

FUTURE WORK

In the future, better methods for parameter optimization will be identified by selecting better parameters that reflects effective knowledge discovery. The role of streaming data processing is still rarely explored when it comes to text classification.

REFERENCES

1. C.Kanakslakshmi, Dr. R.Manickachezian, “An Analysis on Text Mining and Text Classification Techniques”, Proceedings of the National Conference on Information and Image Processing, Volume 1, Page .No 132-135,February 2015.
2. G.Angulakshmi, Dr.R.Manicka Chezian, “ Three Level Feature Extraction For Sentiment Classification”, International Journal of Innovative Research in Computer and Communication Engineering, Volume 2, Issue 8, Page .No5501-5507, August 2014.
3. Amerada Patra, Divakar Singh, “A Survey Report On Text Classification with Different Term Weighing Methods and Comparison between Classifications Algorithms”, International Journal of Computer

Applications, Volume 75, Issue7, Page .No14-18, August 2013.

4. Aakanksha, Er. Dineshkumar, "A Hybrid Approach For Text Classification Using Hmm, Svm And Genetic Algorithm", International Journal For Technological Research In Engineering , Volume 1, Issue 12, Page .No 1454-1457, August-2014.
5. A. Saritha, N. Naveen Kumar, "Effective Classification of Text", International Journal of Computer Trends and Technology (IJCTT), Volume 11,Issue 1, Page .No 1-6, May 2014.
6. J.Sreemathy, P.S.Balamurugan, "An Efficient Text Classification Using Knn And Naive Bayesian", International Journal on Computer Science and Engineering (IJCSE), Volume 4, Issue 3, Page .No 392-396, March 2012.
7. Jafar Ababneh, Omar Almomani,, Wael Hadi, Nidhal Kamel Taha El-Omari, and Ali Al-Ibrahim, "Vector Space Models to Classify Arabic Text", International Journal of Computer Trends and Technology (IJCTT) ,Volume 7 , Issuer 4, Page .No 219-223, January 2014.
8. Megha Gupta, Naveen Aggrawal, , "Classification Techniques Analysis", National Conference on Computational Instrumentation CSIO , Chandigarh, India, Page .No 128-131, March 2010.