Data Harvesting using Machine Learning in Hadoop

Uma S[1], Greeshma G Vijayan[2]

[1]PG student, Dept. of Computer Science and Engineering, LBSITW, Kerala, India
[2]Assistant Professor, Dept. of Computer Science and Engineering, LBSITW, Kerala, India

---***---

Abstract - *The traditional manual classification of news text not only consumes a lot of human and financial resources,but also hardly achieved classification task quality. How to organize the text and makes automatically text classification has become a challenge. It is a challenging task to deal with a large data set, but with the use of Scikit we can easily classify data and gives more accurate results, and then classified by using the classifier. The aim is to create an artificial intelligence program capable of classifying a news article as one of selected categories based on previously experience analyzing training data sets of correctly classified articles. In this Thesis, we collect data sources from online news and classify each data set and gives them a tag for each topic using various machine learning classifier and generate accuracy results. The Goal of this Thesis is to harvest relevant news from online news topics using hadoop and classify them into general categories using various machine learning algorithms. Its accuracy is not as high as human beings but since it can rapidly process data it has its advantages*

***Key Words* : Machine Learning, Artificial Intelligence, Data ming, Text preprocessing.**

1. INTRODUCTION

In today's society a large portion of the worlds population get their news on their electronic devices. Many of the major newspapers have apps that can be used on phones and tablets, in which their articles are displayed in a flow, typically sorted on time of publication. Not all readers are interested in all of the articles published each day. Many newspapers have divided their articles into categories where one can display all articles about e.g. Sports, Fashion or Science. The problem is then that a user would have to go to the different pages to read about different categories. In this paper we explore the possibility of using text classification to teach a machine to select interesting articles based on a users previous experience. With this information a newspaper can create a flow of articles that are personalized for each user.

In modern society,some famous news website such as time server provide information every day for millions of users.But with the continuos development of information technology the amount of disorder data is increasing.Automatic text classification has always been an important application and research topic since the inception of digital documents.Today text classification is a necessity due to the very large amount of text documents that we have to deal with daily.

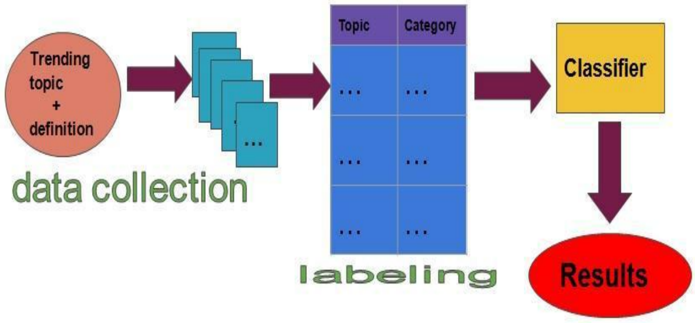
2.MOTIVATION

Since many people use their phones and tablets to access newspaper articles it is easier for users to find new newspapers from all around the world. This leads to a much broader span of customers for the news companies, which makes it more difficult for the companies to keep their customers satisfied. To make this easier they must consider customer care and personalization. We could not find any news companies that offer the kind of personalization presented in this paper. Most have just created pages for different general topics like Sports and Health. The research done in this paper could help newspapers to personalize the flow of articles and provide articles that they know the users wants to read.

3. PROPOSED SYSTEM

Existing model refers that user would have to go to the different websites to read about different categories. In this proposed thesis, we explore the possibility of using text classification to teach a machine to select interesting websites based on a users previous experience

3.1 Data Flow Diagram



The project purpose is to create an artificial intelligence program capable of classifying a news article as one of selected categories based on previous experience analysing training data sets of correctly classified articles.

Basically there are four modules in this project. Firstly, collecting various data articles collection from various web sources. Secondly,Text Preprocessing is the process of preparing and cleaning the data of dataset for classification.It helps to reduce the noise in the text,improve the performance of classifier and speedup the classification process.Tranformation vectorization is supported by various Machine learning algorithms. Here, Multinomial Naive Bayes algorithm is used. The main idea of feature selection is to select a subset of features from the original document. Third module refers to applying classifiers.By Using Multinomial NB machine learning algorithm, datasets are classified and categorised.
Finally,the project is implemented in Big data framework in Hadoop.

3.2 Source Code Algorithm

Step 1: Start
Step 2: Read the dataset.
Step 3: Convert all characters to lowercase letters and remove punctuations
Step 4: Split the data into training and test data
Step 5: Feature extraction and vectorisation
Step 6: Train the model using Multinomial Naive Bayes algorithm
Step 7: Calculate the accuracy score, recall score, precision score and F1 score
Step 8: Stop

3.3 News Article Classification

3.3.1 Load Data

- Records in the file are comma delimited.
- Column titles are included in text file.
- Load the data into a Pandas data frame.
- View unique values for the category column for later transformation to discrete numerical values. Pandas: Python package providing fast, flexible and expressive data structure designed to make working with relational or labelled data.pd is a library variable name.

Lambda: tool for building functions or for building function objects.
Python has two tools for building functions:def and lambda Lambda is a way to create small anonymous functions .i.e functions without name. Mainly used in combination with functions like filter(),map() and reduce().

3.3.2. Preprocess data.

Transform categories into discrete numerical values.
Transform all words to lowercase.
Remove all punctuations.
Bag Of Words-Used in information retrieval. Also known as vector space model.

Limitations of Bag of Words:
Cannot capture phrases and multi-word expressions, effectively disregarding any word order dependence.
Bag of words model doesn't account for potential misspellings or word derivations.

3.3.3. Split into Train and test data sets.

If we split the data ,can apply algorithm or model into one set(training set)and can check result with another set(test data)

3.3.4.Extract features

Apply bag of words processing to the dataset. Feature extraction is very different from feature selection Feature Selection is machine learning technique applied on these features. Feature extraction is transforming arbitrary data(text/images) into numerical features usable for machine learning.
Count Vectorizer: Convert a collection of text documents to a mtrix of token counts. Count vectorizer implements both tokenization and occurrence counting in a single class.

3.3.5. Train multinomial Naive Bayes classifier.

3.3.6. Generate predictions

3.3.7. Evaluate model performance

It is a multi-class classification for evaluating the performance measures such as accuracy score, precision score, recall score score and F1 score.

4.System Architecture

The Data harvesting are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources which includes unstructured and semi structured information. The main goal of text mining is to enable users to extract information from textual resources and to deal with the various operations like, retrieval, Classification (supervised, unsupervised and semi Supervised).Text classification is an important part of Text mining, it is done on the basis of words, phrases And word combinations with respect to set of predefined Class labels.

Text Classification processes consists of training phase and testing phase. In training phase, Dataset is loaded and different classification algorithms are applied to this dataset. After completing the training Phase, performance of classifiers is analyzed and the Classifier which provides the best performance is
Selected. Data mining is useful for extracting or discovering new relation, hidden knowledge and important patterns from huge amount of data. Data
Mining is also known as Knowledge Discovery in Databases (KDD). Data mining uses different technique for knowledge discovery such as classification,
Clustering, summarization, associations etc. Text mining is one of the most important technique used in data mining for analysis of large volume of textual data and it is also one of the key technologies used in data harvesting.

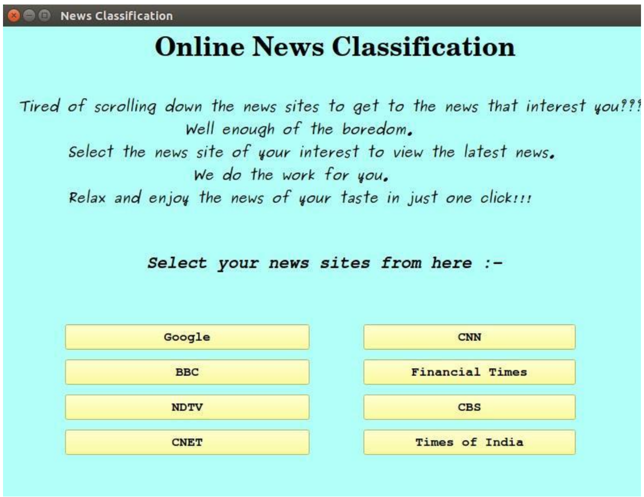
The documents are classified using text classification techniques. This technique is important for categorization of documents in a supervised way. Present research uses text classification technique for classification of news. In this study news data is classified as per the types of news such as
business, sports, entertainment and technology, health, politics, Indian, world, life-leisure and nature. This technique works in two stages. In first stage, it can extract subsequent terms or effective keywords which are useful for identifying class in training phase. In next stage i.e. testing phase actual classification of document is carried out using subsequent terms of keywords. For effectiveness and efficiency purpose these documents are pre-processed. Text is classified using keyword extraction technique. The data is pre-processed by removing stop words which uses stemming, stop words removal and tokenization. Experiments have proved that various classifiers can provide higher accuracy.

5.Results and Discussions

5.1 Module 1

Collecting online news articles live from various websites like google,ndtv etc

- Includes 8 different news article based websites



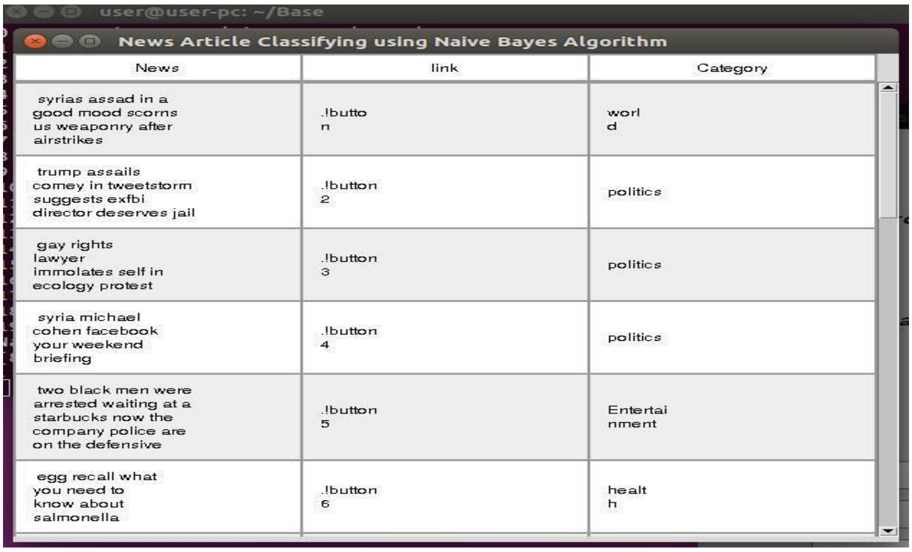
Online News Classification

Tired of scrolling down the news sites to get to the news that interest you???
Well enough of the boredom.
Select the news site of your interest to view the latest news.
We do the work for you.
Relax and enjoy the news of your taste in just one click!!!

Select your news sites from here :-

5.2 Module 2

Classifying each live streamed news article from either website and categorising into specific topic using Multinomial Naive bayes algorithm.

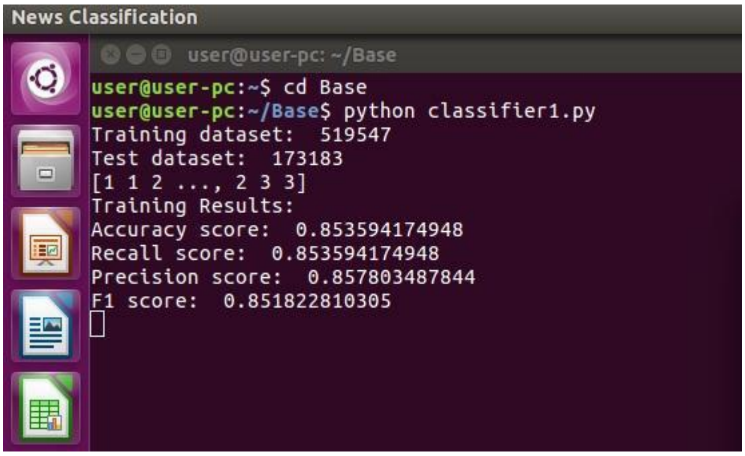


News Article Classifying using Naive Bayes Algorithm

News	link	Category
syrias assad in a good mood scorns us weaponry after airstrikes	.!butto n	worl d
trump assails comey in tweetstorm suggests exfbi director deserves jail	.!button 2	politics
gay rights lawyer immolates self in ecology protest	.!button 3	politics
syria michael cohen facebook your weekend briefing	.!button 4	politics
two black men were arrested waiting at a starbucks now the company police are on the defensive	.!button 5	Entertai nment
egg recall what you need to know about salmonella	.!button 6	healt h

5.3 Module 3

Performance measurement is generally defined as regular measurement of outcomes and results, which generates reliable data on the effectiveness and efficiency of programs. Calculating the Accuracy score,precision,recall and F1 score. Training dataset and testing dataset details are also generated.



```
user@user-pc:~$ cd Base
user@user-pc:~/Base$ python classifier1.py
Training dataset:  519547
Test dataset:  173183
[1 1 2 ..., 2 3 3]
Training Results:
Accuracy score:  0.853594174948
Recall score:  0.853594174948
Precision score:  0.857803487844
F1 score:  0.851822810305
```

6. CONCLUSION

Data harvesting is an extensive area in a news article classification which is discussed in this paper. It has been concluded that results showed that there are different types of categories that has been proposed like politics,financial and sports.entertainment etc. The classification steps i.e data gathering, preprocessing, feature selection and classification algorithm are explained Also the classification process has been implemented using Mutinomial naives bayes classifier. From the result it has been concluded that an accuracy of about 85% has been obtained and has provided good results w.r.t existing methods.

REFERENCES

1. SAMPADA BIRADARa1 And M. M. RAIKAR,"Performance analysis of text classifiers based on news articles"
2. M. IKONOMAKIS,V. TAMPAKAS,"Text
3. Classification Using Machine Learning Techniques",Issue 8, Volume 4, August 2005
4. SHIMA ZOBEIDI, MARJAN NADERAN, SEYED ENAYATOLLAH ALAVI," Effective Text Classification Using Multi-level Fuzzy Neural Network",March 2017
5. F. SEBASTIANI, "Machine learning in automated text categorization," ACM Computing Surveys (CSUR), vol. 34, no. 1, pp. 1–47, 2002.
6. D. B. M. PAZZANI, "Learning and revising user profiles: The identification of interesting web sites," Machine Learning, vol. 27, no. 3, pp. 313–331, 1997
7. ieeexplore.ieee.org/document/8003664/dl.acm.org/citation.cfm?id=1277918
8. Data Mining Concepts and Techniques", Second edition, Jiawei Han and Micheline Kamber.
9. KATHY LEE, DIANA PALSETIA, RAMANATHAN NARAYANAN, MD. MOSTOFA ALI PATWARY, ANKIT AGRAWAL, AND ALOK CHOUDHARYTRUPTI,"Twitter Trending Topic Classification" , 2011 11th IEEE International Conference on Data Mining Workshops
10. ANTONIA KYRIAKOPOULOU and THEODORE KALAMBOUKIS," Using Clustering to Enhance Text Classification"