International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 1387-1391 www.ijaem.net ISSN: 2395-5252

"News Articles Classification"

Mr. Pankaj yadav, Shila Jawale, Mr. Ashutosh Mahadik, Ms. Neha Nivalkar, Dr. S. D. Sawarkar

Datta Megha College of Engineering Airoli, Navi Mumbai Mumbai, Maharashtra
Datta Megha College of Engineering Airoli, Navi Mumbai
Datta Megha College of Engineering Airoli, Navi Mumbai Mumbai, Maharashtra
Datta Megha College of Engineering Airoli, Navi Mumbai Mumbai, Maharashtra
Datta Megha College of Engineering Airoli, Navi Mumbai

Submitted: 01-06-2021 Revised: 14-06-2021 Accepted: 16-06-2021

ABSTRACT— Social media for news consumption is a double-edged sword. On the one hand, its low cost, easy access, and rapid dissemination of information lead people to seek out and consume news from social media. For the last few years, text mining has been gaining significant importance. Since Knowledge is now available to users through variety of sources e.g. electronic media, digital media, print media, and many more. Due to becoming a very hot research area, a lot of unstructured data has been recorded by research experts and have found numerous ways in literature to convert this scattered text into defined structured volume, commonly known as text classification.

Focuses on full text classification e.g. full news, huge documents, long length texts etc. is more prominent as compared to the short length text. We have discussed text classification process, classifiers, and numerous feature extraction methodologies but all in context of texts e.g. news classification based on their headlines. Existing classifiers and their working methodologies are being compared and results are presented effectively.

We also discuss related research areas, open problems, and future research directions for news article classification.

Keywords - News articles, Social Media, Unstructured Data, News Class, News Classification Algorithm.

I. INTRODUCTION

With the rapid growth of online information, text categorization has become one of the key techniques for handling and organizing text data. Text categorization techniques are used to classify news stories, to find interesting information on the World Wide Web and to guide a user's search through hypertext. In these days, most of the available contents are in digital form. To manage such data is big challenge. The textual revolution has seen a tremendous change in the availability of online information. Finding information for just about any need has never been more automatic. Therefore, Text Classification is the task in which sorting is done automatically to classify the documents into predefined classes. Manual text classification is an expensive and time-consuming method, as it become difficult to classify millions of documents manually. Therefore, automatic text classifier is constructed using labeled documents and its accuracy is much better than manual text classification and it is less time consuming too. The proposed work includes the use of Naïve Bayes for online news classification. In the proposed work four types of news has been classified like business, sports, entertainment, political and health. Text classification is the process of assigning text documents to one or more predefined categories. This allows users to find desired information faster by searching only the relevant categories and not the entire information space. To automate the classification process, machine learning methods have been introduced. In a text classification method based on machine learning, classifiers are built (trained)with a set of training documents. The trained classifiers can therefore assign documents to their suitable categories. Online news articles represent a type of web information that are frequently referenced. It will be useful to gather news from these sources and classify them accordingly for ease reference. News Articles classification system, that performs automated news classification. Multinomial Naive Bayes classification method to classify news articles into categories. These categories can be either a set of predefined categories, i.e., general categories, or special categories defined by users themselves. The latter are also known as the personalized categories. With personalized categories, it allows users to quickly locate the desired news articles with minimum effort.

II. PROBLEM DOMAIN

Data mining is the process of sorting through large data sets to identify patterns and

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 1387-1391 www.ijaem.net ISSN: 2395-5252

establish relationships to solve problems through data analysis. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. The data mining process breaks down into five steps. First, organizations collect data and load it into their data warehouses. Next, they store and manage the data, either on in-house servers or the cloud. Data mining programs analyze relationships and patterns in data based on what users request. Data mining techniques are used in many research areas, including mathematics, cybernetics, genetics and marketing. While data mining techniques are a means to drive efficiencies and predict customer behavior, if used correctly, a business can set itself apart from its competition through the use of predictive analysis. Machine learning (ML) is the study of computer algorithms that improve automatically through experience It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

III. TECHNICAL PROBLEM DEFINATION

People need to learn much from texts. But they tend to want to spend less time while doing this. We aim to solve this problem by supplying them the summaries of the text from which they want to gain information. Objective of this project is to detect keywords in the news articles. And categorised them in respective news categories like sport, politics, foods, electronic etc. It will be easier for anyone to search for the particular topics of news and can save much more time for a user. Classification will be done with the help of different Classification algorithms like SVM, Naïve Bayes etc.

IV. EXISTING SYSTEM & ANALYSIS OF THE ISSUES

Text classification system in which they were using unstructured data for their classification which is very difficult to extract because of the large numbers of data which includes media file also so that it is not that easy to extract data from it but the system worked properly in this scenario also but it was more time consuming. Due to the unstructured data it was also not possible to separate the articles or text according to their categories properly this was the another issue faced in this existing system. And the other issue was the user interface it was not that possible for a particular user to search for the respective categories of that particular text document for a news articles.

ANALYSIS OF ISSUES IN EXISTING SYSTEM

- Unstructured data.
- No proper Categories of articles
- No easy to use UI.

Unstructured Data:

Unstructured data in the form of text is everywhere: emails, chats, web pages, social media, support tickets, survey responses, and more. Text can be an extremely rich source of information, but extracting insights from it can be hard and time-consuming due to its unstructured nature. it's often very difficult to analyze unstructured data.

No proper categories of articles:

Due to the unstructured data on internet it is difficult and time consuming process to separate the articles according to the particular categories from that unstructure data(sports, electronics politics etc.) This is one of the issues faced on that system.

No easy to use UI (user interface):

This is one of the issue faced by the user interface used but not that straight forward and quite easy to use classification was done through the user interface was done manually. Due to the manual process of classify the articles it was to time consuming and difficult and also sometimes expensive.

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 1387-1391 www.ijaem.net ISSN: 2395-5252

V. SYSTEM ARCHITECTURE DIAGRAM

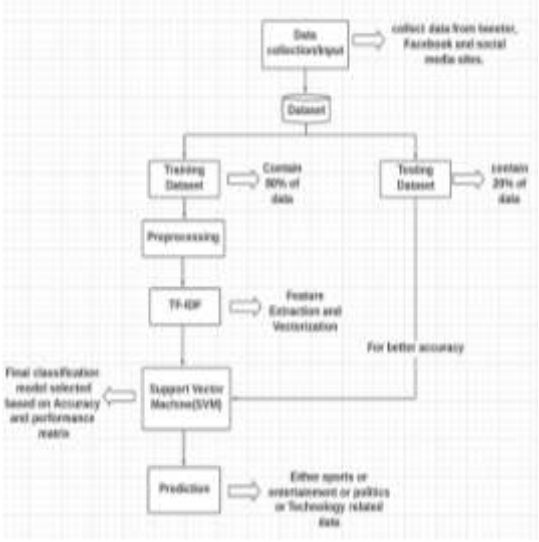


FIGURE: SYSTEM ARCHITECTURE

In the above diagram it show the working of our system. First we collect the the data from different sites like Kaggle.com, collect different data datasets use it has it is or can create a complete new dataset with the use of different datas and form dataset. Second step is to take this dataset and spilt this this dataset into two part training and testing datasets in a ratio of 80:20 where 80% is the train dataset rest is test dataset. Once the dataset is spilt preprocessing method is applied to it and then feature extraction method like TF-IDF to remove unnecessary words after that different classification algorithm are applied on that data to get the best accuracy and performance from the algorithm.
Following are the methods, algorithms used to build this system:

TF-IDF(Term Frequency-Inverse Document Frequency)

There is a large number of terms, words, and phrases in documents that lead to high computational burden for the learning process. Irrelevant and redundant features can hurt the accuracy and performance of the classifiers. Thus, it is best to perform feature reduction to reduce the text feature size and avoid large feature space dimension. There are two different features extraction methods, namely, term frequency (TF) and term frequency-inverted document frequency (TF-IDF). TF-IDF is a weighting metric often used in information retrieval and NLP. It is a statistical metric used to measure how important a term is to a document in a dataset.

Inverse Document Frequency

The inverse document frequency of the word across a set of documents. This means, how common or rare a word is in the entire document set. The closer it is to 0, the more common a word is. This metric can be calculated by taking the total number of documents, dividing it by the number of documents that contain a word, and calculating the logarithm. So, if the word is very common and appears in many documents, this number will approach 0. Otherwise, it will approach 1. Multiplying these two numbers results in the TF-IDF score of a word in a document. The higher the score, the more relevant that word is in that particular document.

Naive Bayes algorithm:

Naive Bayes is a classification algorithm for binary (two-class) and multiclass classification problems. It is called Naive Bayes or idiot Bayes because the calculations of the probabilities for each class are simplified to make their calculations tractable. Rather than attempting to calculate the probabilities of each attribute value, they are assumed to be conditionally independent given the class value. This is a very strong assumption that is most unlikely in real data, i.e. that the attributes do not interact.

Support Vector Machine:

Support Vector Machine (SVM) can be used for regression and classification problems. In regression, SVM predicts a value, whereas in, classification, It is used to predict a class label.SVM is a supervised machine learning algorithm meaning that machine trained with training examples and later trying to predict for new test samples. In SVM, an n-dimensional space is used to plot each data item. Then, a hyper plane which differentiates the classes is used to perform a classification task.

Random Forest:

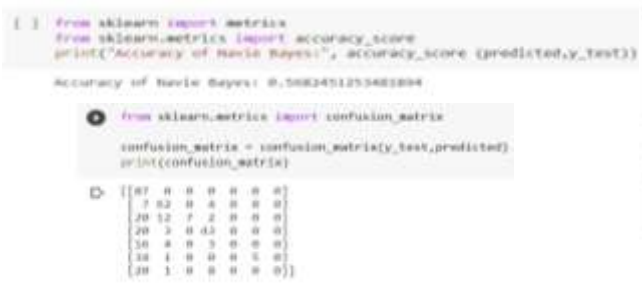
Random forest is a popular machine learning algorithm that belong to the supervised learning technique. It can be used for both classification and regression problems in ML. its is based on the concept of ensemble learning, which is a process of combining multiple classifier to solve a complex problem and to improve the performance of the model. The greater number of trees in the forest lead to higher accuracy and prevents the problem of overfitting.

VI. TESTING AND RESULTS

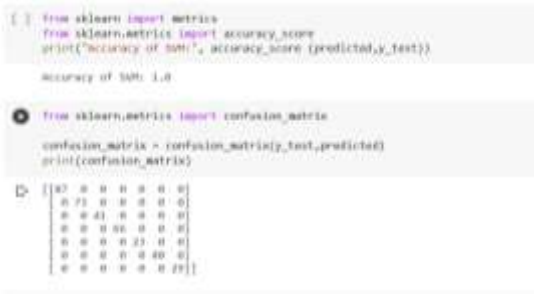
Dataset 1: Accuracy and confusion matrix

Navie Bayes:

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 1387-1391 www.ijaem.net ISSN: 2395-5252



SVM:



Random Forest:

Dataset 2: Accuracy and confusion matrix

Navie Bayes:

SVM:

Random forest:

Result table:

	Dataset No.1	
Sr No.	Algorithms	Accuracy
1	Navie Bayes	56%
2	SVM	99%
3	Random Forest	58%
	Dataset No.2	
Sr No	Algorithms	Accuracy
1	Navie Bayes	92%
2	SVM	94%
3	Random Forest	93%

Above table shows the result of the alogorithm and system applied on the data. We have used two dataset were the first dataset we used was small so the accuracy of the first dataset is not that great compare to second dataset were the accuracy of the model is better. Reason of getting low accuracies is that first dataset have less and not well organized data main reason is data is less in numbers that's the reason we got the less accuracy but SVM give the better accuracy compare to Navie Bayes and Random Forest.

With the other Dataset we got the better accuracy with all three algorithm but in this dataset SVM also performed better with 94% compare to other two algorithms. For this Second dataset

accuracy better because the dataset is well organized numbers of data contain by this dataset is larger compare to first dataset that's the reason accuracy for second dataset is better

VII. CONCULSION

In this paper, we have investigated the possibility to use machine learning algorithm to classify the news articles based on there categories. System shows that problem can be easy solved by using various Classifiers such as Navie Bayes, Support vector machine, Random forest. Support vector machine has out performed the other two classifiers. Random forest has performed well too and Navie Bayes is at the bottom in terms of the performances used in our system. Our future target is to improve the accuracy and also try other classifier like Neural Network.

REFERENCES

[1]. A. K. Durga, and A. Govardhan, September-2011 "Ontology based text categorization – telugu documents", International Journal of Scientific & Engineering Research, Volume 2 Issue 9, ISSN 2229-5518.
[2]. A. McCallum and K. Nigam "A comparison of event models for naive bayes text classification".
[3]. Frank1 and R. R. Bouckaert "naive bayes for text classification with unbalanced.
[4]. V. Gupta and G. S. Lehal (2011), "Punjabi Language Stemmer for nouns and proper name", South and Southeast Asian Natural Language(WSSANLP), IJCNLP, Chiang Mai, Thailand, pp. 35–39.
[5]. http://www.scholarpedia.org/article/Text_categorization
[6]. http://en.wikipedia.org/wiki/Document_classification [6]
[7]. J. D. Brutlag and C. Meek, "Challenges of the email domain for text classification", Microsoft Research, Redmond, WA, 98052 USA.
[8]. K Raghuveer and K. N. Murthy "Text categorization in indian languages using learning approaches" Department of Computer and Information Sciences,University of Hyderabad,Hyderabad.
[9]. Nidhi and V. Gupta, 2012 "Punjabi text classification using Naïve Bayes, Centroid and Hybrid Approach", Sundarapandian et al. (Eds): CoNeCo,WiMo, NLP, pp. 245–252.
[10]. Nidhi and V. Gupta, December-2012 "Domain based classification of punjabi text documents using ontology and hybrid based approach", Proceedings of the 3rd Workshop