Journal of Data Mining and Management
e-ISSN: 2456-9437
Volume 5 Issue 1

Text Data Analysis for Advertisement Recommendation System Using Multi-label Classification of Machine Learning

Rushikesh Chandrakant Konapure*[1], Dr. L.M.R.J. Lobo[2]
[1]PG Student, Department of Computer Science and Engineering, Walchand Institute of Technology, P.A.H. Solapur University, Solapur, Maharashtra, India
[2]Professor, Department of Computer Science and Engineering, Walchand Institute of Technology, P.A.H. Solapur University, Solapur, Maharashtra, India

INFO

Corresponding Author:
E-mail Id: **konapurer@gmail.com*

DOI: *10.5281/zenodo.3600112*

Cite as:
Rushikesh Chandrakant Konapure, & Dr. L.M.R.J. Lobo. (2020). Text Data Analysis for Advertisement Recommendation System Using Multi-label Classification of Machine Learning. Journal of Data Mining and Management, 5(1), 1–6. http://doi.org/10.5281/zenodo.3600112

ABSTRACT

Everyone today can access the streaming content on their mobile phones, laptops very easily and video has been a very important and popular content on the internet. Nowadays, people are making their content and uploading it on the streaming platforms so the size of the video dataset became massive compared to text, audio and image datasets. So, providing advertisements on the video related to the topic of video will help to boost business. In this proposed system the title and description of video will be taken as input to classify the video using a natural language processing text classification method. Aim of Natural Language Processing is to solve the text classification problem by analyzing the contents of text data and decide its category. The proposed system would extract features from videos like title, description, and hashtags based on these extracted features we intend producing classification labels with the use of multi-label classification models. Analyzing produced labels concerning advertisement datasets we intend to provide advertisements on the video related to the topic of the video.

Keywords: *Machine learning, natural language processing, multi-label classification, recommendation system*

INTRODUCTION

Machine learning provides systems the ability to learn and improve from experience automatically without being explicitly programmed. Machine learning is aimed at understanding the data structure and fitting that data into models that people can understand and use. It is possible to categorize most machine learning algorithms into supervised learning and unsupervised learning. Supervised learning algorithms need additional support to train the system using predefined labels. The input dataset is

divided into train and test datasets. The training dataset has an output variable that needs to be predicted or classified. All algorithms learn some kind of patterns from the training dataset and apply them to the test dataset for prediction or classification. In unsupervised learning algorithms, you do not need any extrinsic support. From the provided dataset the learning algorithm tries to find the unknown patterns in data. Clustering and Association methods fall under the category of unsupervised learning algorithms.

MULTI-LABEL CLASSIFICATION

As the availability of data is enormous and it is increasing rapidly so there is an imperative need to organize it. Multi-label classification is an expansion of multiclass classification; multiclass is a single label classification problem which can be categorized into two or more classes. However, in multi-label classification, there is no restriction on the number of labels, one single instance can be assigned to multiple classes.

Multi-label emerged from the analysis of the text categorization problem, where each report may be part of several predefined categories simultaneously A significant problem is the multi-label identification of text and image data. Definitions vary from pieces of media to messages. For example, this can be used to find the genres to which a film belongs based on the summary of the film's plot or poster. In multi-label classification, we need to divide the provided dataset into training and test dataset. Data selected to create a training dataset should not contain any noisy or irrelevant data. The training dataset should contain the elements with appropriate labels so the classifier algorithm can predict the new element's class from the test dataset based on the analysis of the training dataset.

MULTI-CLASS vs. MULTI-LABEL

Dissimilarity in multi-class and multi-label can be explained with the following justification. The components belong to one and the only group in multi-class, while one or more groups can be allocated to each element in multi-label. For example, in the multi-class classification problem, a movie can be classified on the basis of its Censor Board Certification where it belongs to only one class either u/a, r. pg 13 etc., however, in multi-label classification movies can be classified on the basis of the genre the movie belongs to, i.e., it can be a comedy, horror, action, etc. where each movie can be categorized into multiple classes based on the type of movie.

Related Work

Ayon Dey spoke on different machine learning algorithms [1]. The main advantage of using machine learning is that once an algorithm knows what to do with the information, it can do its job automatically. Such algorithms are used to name a few for various purposes such as text classification, spam filtering, object detection, churn prediction, fraud detection, face recognition, etc.

Eva et. al, reviewed the state of the art of multi-label learning and ongoing research [2]. The multi-label model details, as well as the main areas of operation, provided us with the context needed to understand the works examined. This review shows that MLL has been successfully applied to fields such as text, image, video annotation, emotion detection in music, medical diagnosis, prediction of gene and protein activity, and even new areas of application are emerging. They presented an up-to-date overview of multi-label learning for filtering and explaining the main approaches which have been developed so far.

Gangadhara et. al, presented various approaches to solving the classification of

Multi-label [3]. There are many approaches to addressing multi-label labeling issues in a recent study. These are used in different applications such as protein feature classification, music categorization, semantic scene classification, etc. In-turn uses different assessment metrics such as hamming loss and sub-set loss to overcome multi-label classification but which are deterministic? Compared to the binary significance form, this approach's training phase is the same as the binary relevance method's training phase but differs in the testing phase. The second method is based on the idea of related tags being clustered. This method trains one classifier for each group and is called a group representative for the corresponding label. And predict other labels based on the group representative's predicted labels. Based on the concept of association rule mining, the relationships between labels are found.

Jiang et. al, provided a multi-label object categorization approach using convolutional and recurrent neural networks jointly [4]. For the classification of dataset enclosed images with a single label, CNNs are extremely efficient. Traditional approaches to strategies for multi-label object identification perform well but struggle to manipulate the dependencies of the label in an image directly. Recurrent neural networks and Convolutional neural networks are used in jointly to address this issue, the proposed CNN-RNN framework learns a common image-label embedding to characterize semantic label dependence as well as image-label relevance and trained end-to-end from scratch to integrate information into a unified framework.

Kwangsoo et. al suggested a multi-label video identification system on YouTube data [5]. Due to advances in digital technology increase in video data is rapid. Therefore, there is a growing need for techniques to identify moving images automatically. This paper uses NetVLAD and NetFV models as well as the Huber loss function to check the experiment for video classification problems and the YouTube-8 M dataset.

Piotr et. al, introduced scikit-multilearn: a multi-label Python environment based on a scikit [6]. This assortment is firm with the environment of the scikit/scipy and sparse matrices are used for all internal operations. It includes the integration of the most popular algorithms as well as new families of methods such as solutions to the space tag division depending on the network. The problem transformation can help to enhance the adaption rate of deep learning techniques by providing the alternative interface to Keras. This is also followed by other various methodologies like MEKA/WEKA with MULAN parts and enables simple use of these strategies with the rest of the python stack.

Sumit et. al, undertaken a short analysis and outlook of the Brobdingnagian machine learning technologies [7]. Search engines can figure out the information available on the internet with reference to the keywords taken as input from users. It works properly due to the learning algorithm of machine learning. Email spam filters save the consumer from having to buckle down and do many spam email that is additionally associate application for reading. To provide the most sensible perspective of real-world application this paper intends to accumulate each and every technology area under one roof.

METHODOLOGY

The proposed system for Advertisement Recommendation related to the content of the video will be developed as shown in Fig. 1.

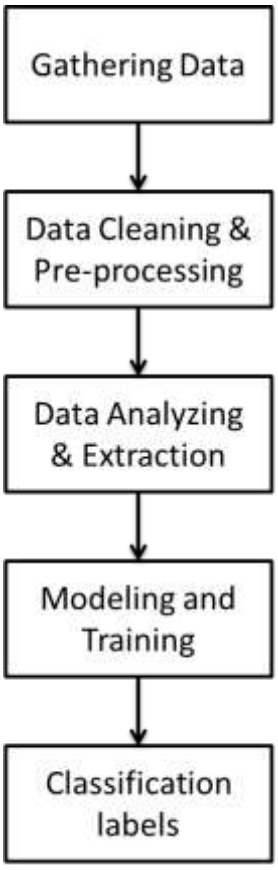

Figure 1: *An overview of the proposed system.*

DATA GATHERING

We get answers to relevant questions in the data collection process and can determine results from an established system by collecting and estimating information on targeted variables. We decided to use YouTube data in the proposed system which is available on the internet. Alternatively, another method is to develop your own database using the YouTube API v3 developed by Google itself for different programmers to communicate with YouTube.

DATA CLEANING AND PRE-PROCESSING

The first step in data pre-processing is to handle the missing data. Since missing values are supposed to be text data in the proposed system there is no way to substitute them, thus the only option is to remove them. Following this on remaining data we will be performing natural language processing text cleaning techniques to get clean and required data. This approach is broken down into the following steps:

Converting to Lowercase

It is a good practice to keep all text in the same format, however, converting capital words into lowercase does not change the meaning of word Eg. 'Football' and 'football' are semantically the same.

Removing numerical values and punctuation

Numerical values and special characters used in punctuation ($,! etc.) do not contribute to determining the correct class.

Removing extra white spaces

Such that each word is separated by a single white space, else there might be problems during tokenization.

Tokenizing into words

This refers to splitting a text string into a

list of ‘tokens’, where each token is a word. For example, the sentence ‘He sat under a tree’ will be converted to [‘He’, ‘sat’, ‘under’, ‘a’, ‘tree’].

Removing non-alphabetical words and ‘Stop words

Stop words refer to words like and, the, is, etc., which are important words when learning how to construct sentences but of no use to us for predictive analytics.

Lemmatization

In Lemmatization, we convert a word into its base meaning. For example, the words ‘Playing’ and ‘played’ will be converted to their base format ‘play’.

DATA ANALYZING AND EXTRACTION

Data extraction and data analysis are two different techniques where data cleaning, transformation, and data modeling come under analysis part to discover useful information and to support decision-making. And the data extraction process consists of retrieval of information from the available data sources for further data processing. As machines do not understand our language we need to convert the available data into machine-readable language. That is, the text data should be converted into numerical based features so that the computer can build a mathematical model as a solution. In the proposed system we expect the label as output which is in the categorical format, so we need to assign a number to each category. Using a feature vector we can portray our text data into the numerical format. The vectorization method can be explained with the following example of a fruit classification problem in which the feature vector contains the color, size, and number of seeds in a numerical format based on which we label fruit. We plan to use the TFIDF vectorizer, which is often used in text mining to know the value of a word in a text as well as to turn the text into meaningful information within the proposed system.

MODELING AND TRAINING

Data modeling is used to define and analyze the data needed to support the machine learning model and store it into a database to perform required operations on it. In data modeling, the processed dataset will be divided into a training dataset and a test dataset. In the proposed system we from a training dataset with the features extracted using the TFIDF vectorization method. We have a YouTube dataset containing the title and description we check whether our extracted features make any sense on the basis of finding out the most correlated words. We form our training set with the best keywords for each class using title features and the same with the description features where the selected keywords are correlated to the corresponding class. Here, the class is categories of videos like 'science and technology', 'travel', and the keywords most related to these categories are respectively as science, computer, technology, quantum, blogger, vlog and trip, etc. Training datasets are used to train the machine learning algorithm to understand the data to predict the labels of that data the model sees and learns from data. Whereas, testing data is only used once the model is completely trained to evaluate the model in terms of its performance and accuracy. The test set should contain sample data extents the various classes our model would face when used in real-world datasets. The division ratio of training and test dataset did on the basis of the total number of samples in a dataset and the actual model we are training.

CLASSIFICATION LABELS

It is a process of producing labels i.e. output from the classifier model. The labels will be related to the content so from labels, we can understand what the content is about.

EXPECTED RESULT

The main objective of this project is to extract features from videos and produce labels related to the content of a video to recommend the advertisements related to the content using a multi-label classification of machine learning.

CONCLUSION

According to many users, they get annoyed by the irrelevant advertisements displayed while streaming online so they have to wait till completion of video to avoid it. However, with this system, we can boost business by accurate online advertising which is related to the content by detecting the topic of the video. For instance, we can put some company who provides web hosting on website development tutorial videos rather than any random video. Therefore, the client can distribute their advertisements more purposely and accurately to that topic, meanwhile, the user will not be annoyed by the advertisement and at the same time, they might watch the whole video or visit the client's website for more information which will attract more customers.

REFERENCES

1. Ayon Dey (2016), "Machine learning algorithms - A Review", *International Journal of Computer Science and Information Technologies*, Volume 7, Issue 3, pp. 1174-1179.
2. Eva Gibaja, Sabastian Ventura. (November 2014) "Multi-label learning: A review of the state of the art and ongoing research", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*., pp.1-46.
3. Gangadhara Rao Kommu, M.Trupthi, Suresh Pabboju (1-2 August 2014), "A novel approach for multi-label classification using probabilistic classifiers", *IEEE International Conference on Advances in Engineering & Technology Research (ICAETR)*, Unnao, India.
4. Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang Wei Xu1, (2016), "CNN-RNN: A unified framework for multi-label image classification", *Computer Vision and Pattern Recognition*, pp. 1-10.
5. Kwangsoo Shin, Junhyeong Jeon, Seungbin Lee, Boyoung Lim, Minsoo Jeong, Jongho Nang (2016), "Approach for video classification with multi-label on YouTube-8M dataset", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5297-5307.
6. Piotr Szymanski, Tomasz Kajdanowicz (2018), "A scikit-based Python environment for performing multi-label classification" *Journal of Machine Learning Research*, Volume 5, pp.1-22.
7. Sumit Das, Aritra Dey, Akash Pal, Nabamita Roy (April 2015), "Applications of artificial intelligence in machine learning: Review and Prospect" *International Journal of Computer Applications*, Volume 115, Issue 9, pp. 31-41.