INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

Trustworthiness Valuation of Web data Contents using machine learning classifier

[1]SINGAM PRABHAKAR REDDY, **[2]M. SAMUEL SANDEEP REDDY**

[1]PG Scholar, Dept. of MCA, Newton's Institute of Engineering, Guntur, (A.P)

[2]Associate professor, Dept. of CSE, Newton's Institute of Engineering, Guntur, (A.P)

Abstract: Most students rely on online courses as their second-most-used learning strategy, just behind traditional classroom instruction. The development of the internet and e-learning technologies has led to a meteoric rise in the dissemination of knowledge through these mediums. Without e-learning tools, delivering knowledge to certain individuals would have been impossible. Most working professionals engage in concentrated study to better themselves professionally, whether via promotion or just expanding their area of expertise. A plethora of resources exist for these students. Most students rely on online courses as their second-most-used learning strategy, just behind traditional classroom instruction. The development of the internet and e-learning technologies has led to a meteoric rise in the dissemination of knowledge through these mediums. Without e-learning tools, delivering knowledge to certain individuals would have been impossible. Most working professionals engage in concentrated study to better themselves professionally, whether via promotion or just expanding their area of expertise. In their areas of interest, these students may quickly and readily locate several free e-learning web sites on the Internet. However, it might be challenging to locate the most appropriate e-learning material for their learning depending on their current degree of subject expertise. Users wasted their time trying to find useful information from the mountain of accessible data.

Keywords: **Bag of Word, Computer Science, Scholarship, Convolutional Neural Network, recipients, E-Learning.**

I INTRODUCTION

The term "e-learning" refers to the use of the internet and computers to facilitate education. Learning that is supported by technology bridges the physical distance between student and instructor and makes lifelong education possible. The proliferation of e-learning platforms and resources has contributed much to the industry's meteoric rise in recent years.

Many business owners and professionals have recognised the value of e-learning and are taking use of it in their daily lives. With the advent of e-learning, students are able to study on their own time and schedule, from any location of their choosing. Nowadays, the e-learner may study anywhere, as long as they have access to a computer, tablet, or smart phone. Students will no longer be limited by factors like location and schedule[1]. There has been a meteoric rise in the number of e-learning websites and related technologies as e-learning has gained in popularity. Without the advent of e-learning, it would have been impossible to impart knowledge to some individuals. Learners can easily find multiple e-learning web sites for their domain, whether it be medicine, engineering, science, or the social sciences, and this opens up the possibility of continuous education through e-learning. E-learning has the potential to revolutionise the way information is disseminated and to radically alter the educational experience of its participants. However, the human aspect is crucial to learning and is absent in all e-learning technology. It hinders communication and gets in the way of the best route to education. Since this is the case, e-learning's efficacy and outcomes cannot be guaranteed.

II LITERATURE SURVEY

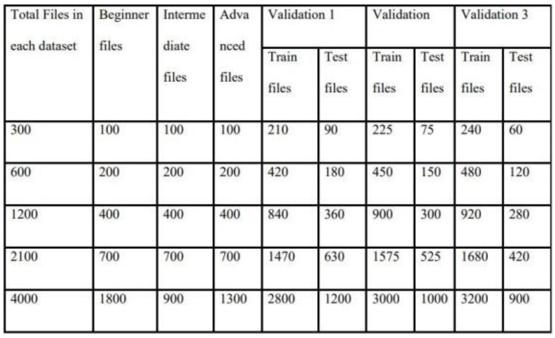
Recommendation categorization utilising social and content-based information was suggested by C.Basu et al. It predicts the user's preferences and tastes based on the customer rating and other information about the product. When compared to the current social-filtering strategy, the model provides superior movie suggestions[37]. M. Claypoolet al. presented a recommendation model that weighted-average Content-based and Collaborative filtering. The methodology makes use of both the depth and speed of content-based filtering to take use of expert opinion and human intelligence. The model is tested on a corpus of recently published articles available online, and positive results are reported[38]. As the computational complexity of user-based collaborative filtering systems grows linearly, George Karpas and Mukund D suggested an item-based top N recommendation method to solve scalability problems. The user-item matrix is analysed using the item-based model approach to uncover hidden relationships between things. The approach determines the degree of similarity between the user's current shopping basket and the suggested item by analysing the degree of similarity between each item in the basket. Experiments demonstrated a threefold increase in speed compared to standard collaborative recommendation methods[39]. A probabilistic relational model for collaborative filtering was presented by Lise Getoor and M. Sahami. This approach allows aspects of things to rely probabilistically on other qualities of objects, which helps describe the uncertainty of connection existence. The model is a crucial part of collaborative filtering since it gives a basic representation of complicated entity interaction. The model substitutes a graph probabilistic structure for the Bayesian Network, which is more intuitive. An approach suggested by G.Nathaniel et al. combines collaborative filtering with personal agents to improve suggestion accuracy. This study demonstrated that combining the efforts of individual agents with those of a community may provide superior recommendation outcomes than those achieved by employing each factor alone. Collaborative filtering, when used to build a team of autonomous agents, may outperform other combination strategies. The collaborative filtering framework chooses the agents for the user automatically. The framework's results corroborated the notion that a group of individuals with a shared perspective may make better recommendations than a single expert could [41]. J. L. Herlocker et al. presented a user-preferences-driven, task-based recommendation system. When making recommendations, most systems instead prioritise existing ratings data above user preferences. It all depends on the work at hand or the environment in which the activity is being performed. The suggested method is based on task-based recommendations, which are not affected by the substance of the situation. The system must do task profiling in the task-focused method to determine the nature of the work the user intends to perform. The user's description of the job is used to create a ranking list that will be used to provide recommendations to the user. A series of helpful suggestions is provided for the user. The same is put through its paces using real-world examples of suggested systems, and the findings are verified empirically[42].

III. METHODOLOGIES

The E-contents are collected from multiple e-learning websites such W3schools, geeks for geeks, rose India, java world, wide skills, oracle docs, java2s, java point, guru99, study tonight, Eureka, code side, journaled, beginner book, tutorial point, tutorial ride, tutorial cup, java2blog, merit campus, data-flair-training, IBM developer. For this purpose, scrappy. spider is used to crawl the sites. The Spider explores the website, downloads its content, and extracts text files using the beautiful soup parser API, all of which are then saved to the local machine. To evaluate the performance of the classifier on a wide range of data sizes, many data sets are generated.

The framework is tested with text files of varying sizes (300 to 4000 documents) containing varying amounts of data. There are three levels of difficulty for each dataset: basic, moderate, and advanced. These directories include data collected across three different degrees of challenge. The tabular description of the dataset’s structure follows.

Table 3:1 Description of Dataset obtained from websites through web scrapping.



Total Files in each dataset	Beginner files	Intermediate files	Advanced files	Validation 1		Validation		Validation 3	
				Train files	Test files	Train files	Test files	Train files	Test files
300	100	100	100	210	90	225	75	240	60
600	200	200	200	420	180	450	150	480	120
1200	400	400	400	840	360	900	300	920	280
2100	700	700	700	1470	630	1575	525	1680	420
4000	1800	900	1300	2800	1200	3000	1000	3200	900

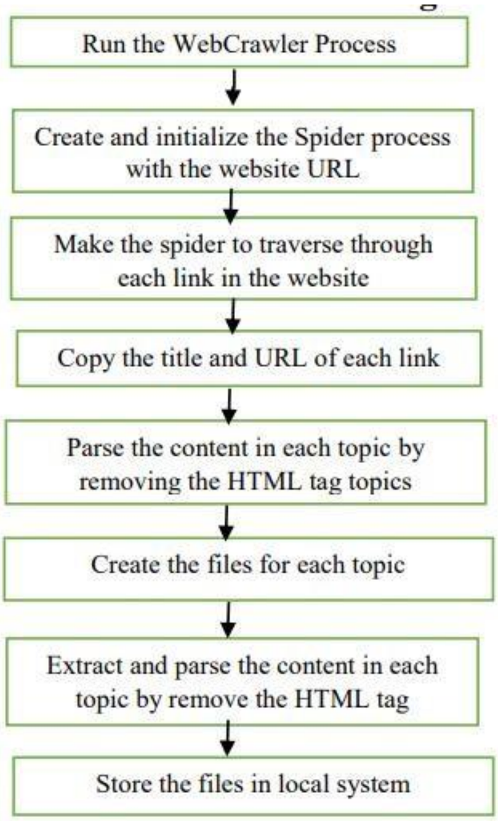

Run the WebCrawler Process
Create and initialize the Spider process with the website URL
Make the spider to traverse through each link in the website
Copy the title and URL of each link
Parse the content in each topic by removing the HTML tag topics
Create the files for each topic
Extract and parse the content in each topic by remove the HTML tag
Store the files in local system

Figure 3:1 Work flow of the web crawler used for data collection

In the diagram, you can see how the web crawler worked to gather the data for the study.

The WebCrawler is a spider that is set up using the site's URL as input. The spider follows every link on the website, collecting the URL and the title as it goes. The material is downloaded, and a file is made with the title and subject for each topic.

text content was written into the file after HTML tags were parsed out, decomposed, and discarded. Put the document in a local folder. All of the website's links are saved as text files. The dataset consists of these text files.

IV RESULT ANALYSIS

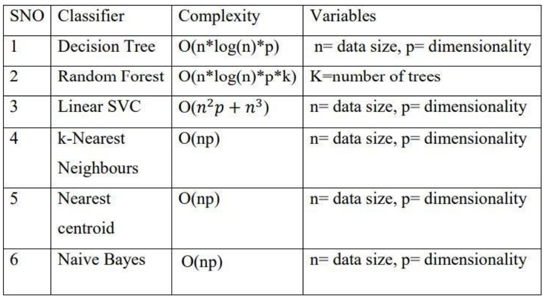
The resource conception or computational cost is used to determine the computational efficiency. The time and storage space needed to execute the procedure are taken into account. A comparison of time and space complexity for gauging processing speed and storage capacity.

Analyse the computational complexity of an algorithm by considering the time and memory it would take to execute the method. Since the number of resources required changes with the size of the input value, complexity is expressed as a function of n, where n is the size of the input value.

Time complexity and space complexity are used to evaluate the algorithm's effectiveness. Time complexity is a measure of how long it takes to execute an algorithm.

algorithm. Time complexity is usually estimated by counting the number of elementary operations performed by the algorithm, supposing that each elementary operation takes a fixed amount of time to perform. Space complexity is the quantity of memory needed to execute an algorithm in relation to its input data size.

Table 4:4: The time and space complexity calculated using the functions.



SNO	Classifier	Complexity	Variables
1	Decision Tree	O(n*log(n)*p)	n= data size, p= dimensionality
2	Random Forest	O(n*log(n)*p*k)	K=number of trees
3	Linear SVC	O(n^2p+n^3)	n= data size, p= dimensionality
4	k-Nearest Neighbours	O(np)	n= data size, p= dimensionality
5	Nearest centroid	O(np)	n= data size, p= dimensionality
6	Naive Bayes	O(np)	n= data size, p= dimensionality

Time complexity for n data size and P dimensions is shown in Table 4:4 for a variety of text classification techniques. Real-world algorithms have a complexity of O(np), where n and p are random numbers and are determined by regression analysis of randomly produced samples.

The suggested approach employs a Complexity evaluator object to quantify the time and space requirements. Parameters for the Complexity evaluator objects include content pathways, a feature list, and the names of classifiers.

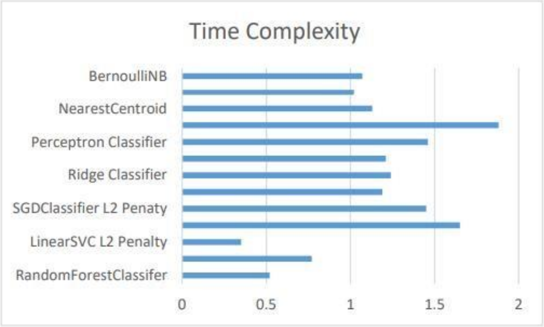


Figure 4:1: Time complexity of different classifiers on the dataset

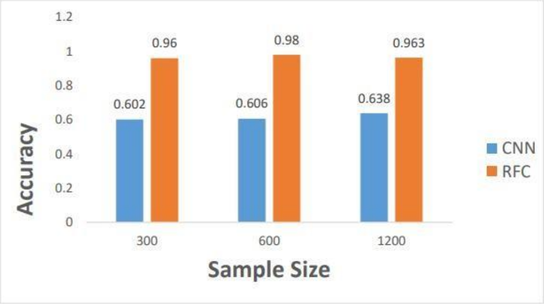


Figure 4:2: Comparison of Random Forest with CNN classifier.

demonstrates how Convolutional Neural Networks and the Random Forest classifier are pitted against one another to verify the efficacy of deep learning.

The random forest consistently beat the CNN in the tests. The size of the dataset and the number of documents within it are two crucial factors that are taken into account throughout the comparison. Several types of information are used to evaluate the executed outcomes. The findings demonstrated a significant gap in precision between CNN and RFC. When compared to CNN, RFC consistently produces higher accuracy over a wide range of sample sizes.

V CONCLUSION

Because of developments in the internet and e-learning websites, both the quantity of e-learning websites and the number of people using them have exploded in recent years. There is so much e-content out there that it might be daunting for the user to choose the correct material.

As a result, it's crucial to have a smart way of determining what e-learning materials the user needs. The proposed system trains a model using a Random Forest Classifier. The approach is used to the task of determining how tough certain pieces of online learning material are. With this system, users may narrow down a vast online library to relevant learning resources that are tailored to their specific skill set.

Web scraping is used to get the data, which is then included into the framework. Machine learning document classification techniques are used to the dataset after pre-processing and feature selection in order to determine the most effective classification scheme. In order to train the framework, the Random Forest classifier is used. The framework recommends e-contents from different E-Learning websites, categorising them as either appropriate for a beginning, an intermediate learner, or an advanced learner.

REFERENCES

[1] U. Dw et al., "The_role_of_m-learning_in_the_future_of_elearning_in_Africa.pdf (Object to application/pdf)," pp. 1–12, 2003.

[2] R. Mahajan, P. Gupta, and T. Singh, "Massive Open Online Courses: Concept and Implications," Indian Pediatr., vol. 56, no. 6, pp. 489–495, 2019.

[3] E. El Bachari, E. H. Abelwahed, and M. El Adnani, "E-Learning personalization based on Dynamic learners' preference," Int. J. Computer. Sci. Inf. Technol., vol. 3, no. 3, pp. 200–216, 2011.

[4] N. M.-C. & M. Fuertes-Aplite., E-Learning Research Report 2017. 2017.

[5] G. Geetha, M. Safa, C. Fancy, and D. Saranya, "A Hybrid Approach using Collaborative filtering and Content based Filtering for Recommender System," J. Phys. Conf. Ser., vol. 1000, no. 1, 2018.

[6] J. K. Tarus, Z. Niu, and G. Mustafa, "Knowledge-based recommendation: a review of ontology-based recommender systems for e-learning," Artif. Intell. Rev., vol. 50, no. 1, pp. 21–48, 2018.

[7] K. I. Ghauth and N. A. Abdullah, "Learning materials recommendation using good learners' ratings and content-based filtering," Educ. Technol. Res. Dev., vol. 58, no. 6, pp. 711–727, 2010.

[8] S. Pariserum Perumal, G. Sannasi, and K. Amruthraj, "An intelligent fuzzy rule-based e-learning recommendation system for dynamic user interests," J. Supercomputer., no. 0123456789, 2019.

[9] P. Viana and M. Soares, "A Hybrid Approach for Personalized News Recommendation in a Mobility Scenario Using Long-Short User Interest," Int. J. Artif. Intell. Tools, vol. 26, no. 2, pp. 1–25, 2017.

[10] N. Srivani, Dr Prasadu Peddi, "Efficient Fr a Geometrical-Model-Based Face Segmentation and Identification in Terms of Identification the Face ", *JFCR*, pp. 1283-1295, Jun. 2022.

[11] Z. Piao, S. J. Yoo, Y. H. Gu, J. No, Z. Jiang, and H. Yin, "Recommender system architecture based on Mahout and a main

memory database," J. Supercomputer., vol. 74, no. 1, pp. 105–121, 2018.

[12] Prasadu Peddi (2023), Using a Wide Range of Residuals Densely, a Deep Learning Approach to the Detection of Abnormal Driving Behaviour in Videos, ADVANCED INFORMATION TECHNOLOGY JOURNAL, ISSN 1879-8136, volume XV, issue II, pp 11-18.

[13] Y. Wang and W. Shang, "Personalized news recommendation based on consumers' click behaviour," 2015 12th Int. Conf. Fuzzy Syst. Knowl. Discov. FSKD 2015, pp. 634–638, 2016.

[14] Naga Lakshmi Somu, Prasadu Peddi (2021), An Analysis Of Edge-Cloud Computing Networks For Computation Offloading, Webology (ISSN: 1735-188X), Volume 18, Number 6, pp 7983-7994.