An Improved Approach for Fast Documents Scrapping and Classifying Using Selenium Automation and Multinomial Naïve Bayes Classifier

[1]Sunny Mehta, [2]Gayatri Jain (Pandi)

[1]Master of Engineering Student, [2]Head of Department
[1]Computer Engineering,
[1]L.J.Institute of Engineering and Technology, Ahmedabad, India.

Abstract: Generally, Selenium Automation is used for testing purpose and detecting errors and defects of the system in development. However, we will use Selenium for making a list of required web element from web page. In addition, by using that list selenium will identify the new documents from web pages for scrapping data. For example, we consider any tenders site, there may be thousands of tenders getting published every day, so it is very hard for the user to surf every tender one after another to get the tender of his/her need. But, in our method firstly, we are using bag of words method to gather test data for further classification. Secondly, we are using Multinomial Naïve Bayes Classifier to classify our documents industry wise which will be useful for the user to pick up his category fresh tender. For picking up a fresh tender, user will access the folder created on the desktop where the scraped fresh documents will be stored in a technology wise folder. In the last, Confusion Matrix will be built and detailed accuracy by class for the technology category will be calculated and shown. This approach helps the larger service providing business organizations to provide their clients the documents of their needed categories.

Index Terms **– Selenium Web Driver, Multinomial Naïve Bayes Classifier, Bag of Words, Stream Writer, Web Scrapping.**

I. INTRODUCTION

1.1 Selenium

Selenium is a portable framework for testing web apps. It also provides a Selenese - a specific language to write tests in a number of popular programming languages, including C#, Groovy, Java, Perl, PHP, etc. The tests can then run on most web browsers. It runs on Windows, Linux, and macOS. It is open source software released under the Apache License 2.0 [1].

Fig-1.1: Modern browsers use to run tests of Selenium Framework.

1.2 Difference between Naïve Bayes and Multinomial Naïve Bayes Classifier

Naive Bayes classifier is the term which refers to conditional independence of each of the features in the model, while Multinomial Naive Bayes classifier is the specific instance of a Naive Bayes classifier which uses a multinomial distribution for each of the features [2].

It's a family of probabilistic algorithms based on applying Bayes theorem with the naive assumption of conditional independence between every pair of a feature. Bayes theorem calculates probability P(c | x) where c is the class of the possible outcomes and x is the given instance which has to be classified, representing some of the certain features [3].

P (c | x) = P (x | c) * P(c) / P(x)

1.3 Difference Between Multivariate Bernoulli and Multinomial Naïve Bayes

Multinomial cares about the counts for a multiple feature that do occur, whereas Bernoulli cares about the counts for a single feature that do occur and counts for the same feature that do not occur.

It means that, for example, Multinomial will classify a document based on the counts it finds of the multiple keywords. Whereas, Bernoulli can only focus on a single keyword but, will also count how many times that keyword does not occur in the document.

However, they do model slightly different things. If you have discrete multiple features to worry about, you have to use Multinomial Naïve Bayes Classifier.

1.4 Bag of Words

The bag of words model is a simplifying representation used in natural language processing and information retrieval. In this model, it represents text as the bag of its words, disregarding grammar and even word order but keeping multiplicity. In addition, it has also been used for computer vision.

Generally, this model is commonly used in methods of document classification where the occurrence of each word is used as a feature for training a classifier. [4]

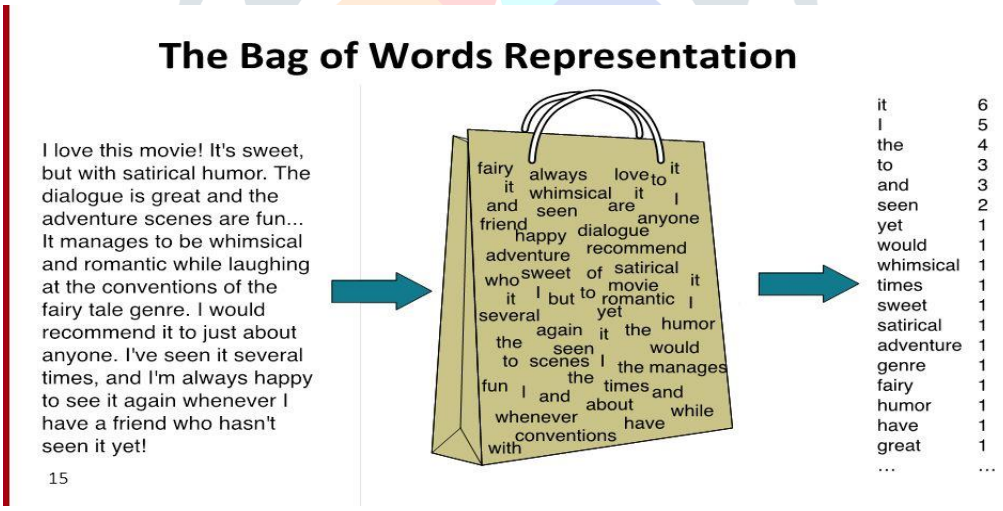

Fig-1.2: Bag of Words Representation.

II. RELATED WORK

Generally, in recent trends in selenium is for the better automated testing purpose and for efficient test results rather than doing a manual testing which is more time consuming and human effort is needed. System automatically learns what actions to be done where on its own. For example, if there is a button it will click there or else if there is a text box then it will enter the required text etc. [6]. System automatically learns what actions to be done where on its own [7]. For example, if there is a button it will click there or else if there is a text box then it will enter the required text etc. The quality of web application is one of important factor while deploying the web applications [9].

By using Selenium automation framework test scripts system will automatically finds the mentioned elements from the web pages where the action is mentioned and whenever the element is not found then it will inform the user that this element is not found in the web page so it will be easy for the user to just change the script for that element and the test script will run automatically from the next time [8].

Since there is a consistent growth in the volume of digital documents, both on the internet and within organizations, the need to classify them into categories is obvious. So, bag of words technique to represent the tender documents. Classification was implemented by Naïve Bayes classifier [5].

III. PROPOSED MODEL

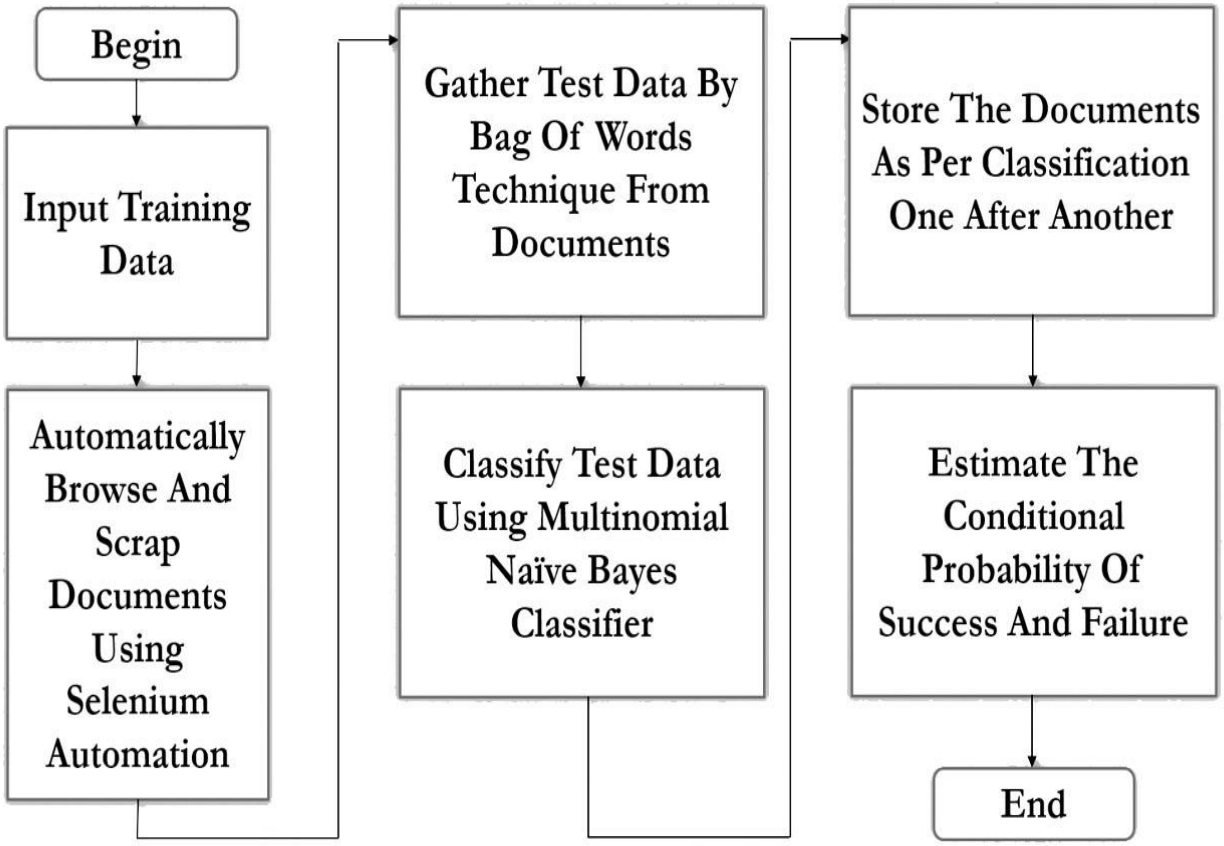

Fig-3.1: Proposed Model Flowchart.

Figure 3.1 illustrates the working principles of this paper proposed system. Firstly, the system starts with preprocessing the data where the website is automatically simply scrapped documents. After that, text from every document is extracted to create the instances.

Secondly, whole website is started scrapping by selenium automation framework for classifying documents in the respective technological categories.

Thirdly, by using bag of words approach gathering the text for test data and afterwards that data is used in classification of documents in their technological category by using Multinomial Naïve Bayes classifier. In addition, documents will be stored in technologically classified folders.

At last, system will calculate the detail accuracy by class for technological category.

IV. IMPLEMENTATION

The proposed approach mainly focuses on improving the accuracy for classification based on technological category. The detailed accuracy by class for technology category is calculated based on the following formulas.

1.) TP Rate = TP / (TP + FN)
2.) FP Rate = FP / (FP + TN)
3.) Precision = TP / (TP + FP)
4.) Recall = TP / (TP + FN)
5.) F1 Score = 2 * ((Precision * Recall) / (Precision + Recall))

The proposed method yields the accuracy of 81.481% using which is more than the existing system accuracy using Multinomial Naïve Bayes classifier. The detailed step wise implementation with its snapshots are shown below.

4.1 Preprocessing

Firstly, using Selenium Automation all the DRDO documents from the website are automatically downloaded and from that a list of words was prepared for each file and the unwanted words (and, or, the, but, …) were further removed. A list of 1,49,245 unique words was then built for classification after removing the duplicates from the collective word list of all the documents.

Keyword	Product_Name	SubIndustry_Name	Industry_Name
3D Printer	Abrasive And Abrasive Media	Abrasives And Abrasive Media	Machinery And Tools
Delivery Of Manhole Covers	Abrasive And Abrasive Media	Abrasives And Abrasive Media	Machinery And Tools
Lighting Panel Scrap	Abrasive And Abrasive Media	Abrasives And Abrasive Media	Machinery And Tools
Revival To Pws Scheme At	Abrasive And Abrasive Media	Abrasives And Abrasive Media	Machinery And Tools
3D Scanning Systems	Abrasive Wheels	Abrasives And Abrasive Media	Machinery And Tools
Delivery Of She Toilets	Abrasive Wheels	Abrasives And Abrasive Media	Machinery And Tools
Lighting Poles	Abrasive Wheels	Abrasives And Abrasive Media	Machinery And Tools
Revolvers	Abrasive Wheels	Abrasives And Abrasive Media	Machinery And Tools
2D Image Scanners	Aadhaar Enabled Payment System	Accounting Software	It And Telecommunications
Dehydration Units	Aadhaar Enabled Payment System	Accounting Software	It And Telecommunications
Lighting equipment and electric lamps	Aadhaar Enabled Payment System	Accounting Software	It And Telecommunications
Revival Of Mws Scheme	Aadhaar Enabled Payment System	Accounting Software	It And Telecommunications
Abrasion Powder	Accounting Software	Accounting Software	It And Telecommunications
Dental Equipments and Stores	Accounting Software	Accounting Software	It And Telecommunications
Lime Preparation Plants	Accounting Software	Accounting Software	It And Telecommunications
Rhinoplasty Sets	Accounting Software	Accounting Software	It And Telecommunications
Abortion Equipment	Accounting Software	Accounting Software	It And Telecommunications
dental equipment	Accounting Software	Accounting Software	It And Telecommunications
Lime Kiln Repair	Accounting Software	Accounting Software	It And Telecommunications
Rhinoplasty Instrument	Accounting Software	Accounting Software	It And Telecommunications
Abrasion Powders	Accounting System	Accounting Software	It And Telecommunications
dental equipments	Accounting System	Accounting Software	It And Telecommunications
Lime Quick	Accounting System	Accounting Software	It And Telecommunications
Rhinoplasty Sets	Accounting System	Accounting Software	It And Telecommunications
AC Control Panel Repairs	Active Tracker	Accounting Software	It And Telecommunications
Dental Workstations	Active Tracker	Accounting Software	It And Telecommunications
Linear Motion Guide	Active Tracker	Accounting Software	It And Telecommunications
Ridge Planterss	Active Tracker	Accounting Software	It And Telecommunications
Agricuture Equipmentsss	Application Software	Accounting Software	It And Telecommunications
Development Of Streetssss	Application Software	Accounting Software	It And Telecommunications
Lot Name Kv Vcbss	Application Software	Accounting Software	It And Telecommunications
Roof Maintenance	Application Software	Accounting Software	It And Telecommunications
Aids Kit	Archiving Service	Accounting Software	It And Telecommunications
development of wall ss	Archiving Service	Accounting Software	It And Telecommunications
Lot Name Mis Wheel Discss	Archiving Service	Accounting Software	It And Telecommunications

Fig-4.1: Dataset Sample.

4.2 Chrome browser is controlled by automated testing software

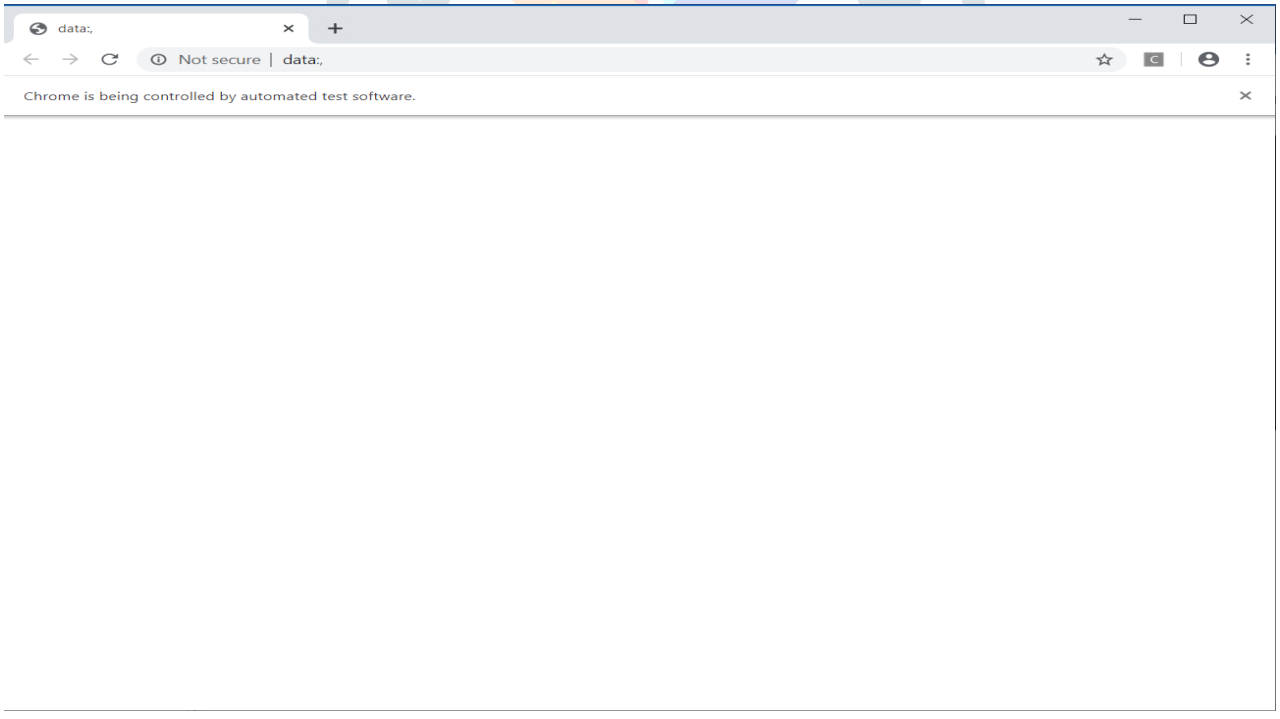

Fig-4.2: Chrome browser is controlled by automated testing software.

4.3 Gathering text from documents using Bag of words approach for classification

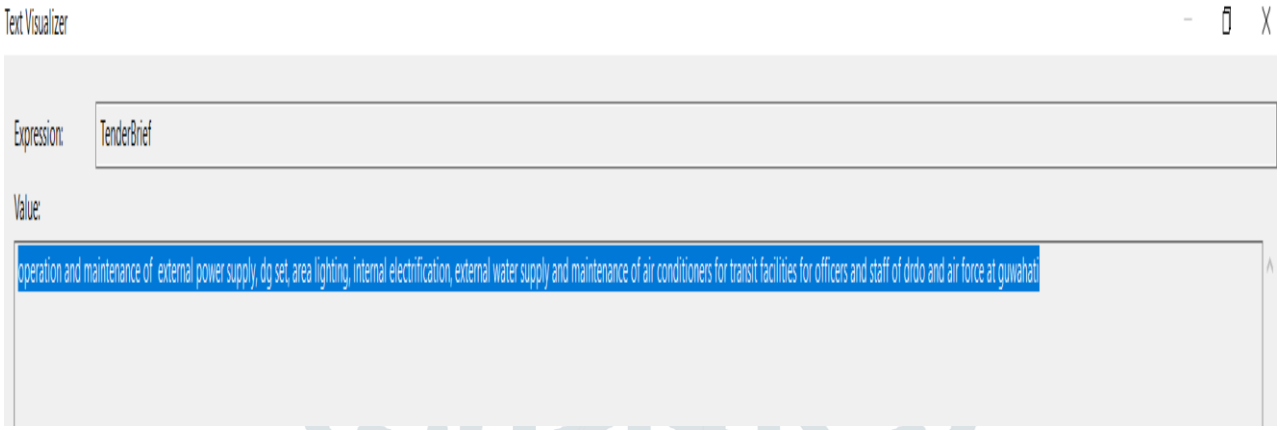

Fig-4.3: Gathering text from documents using Bag of words approach for classification.

4.4 Proposed system output

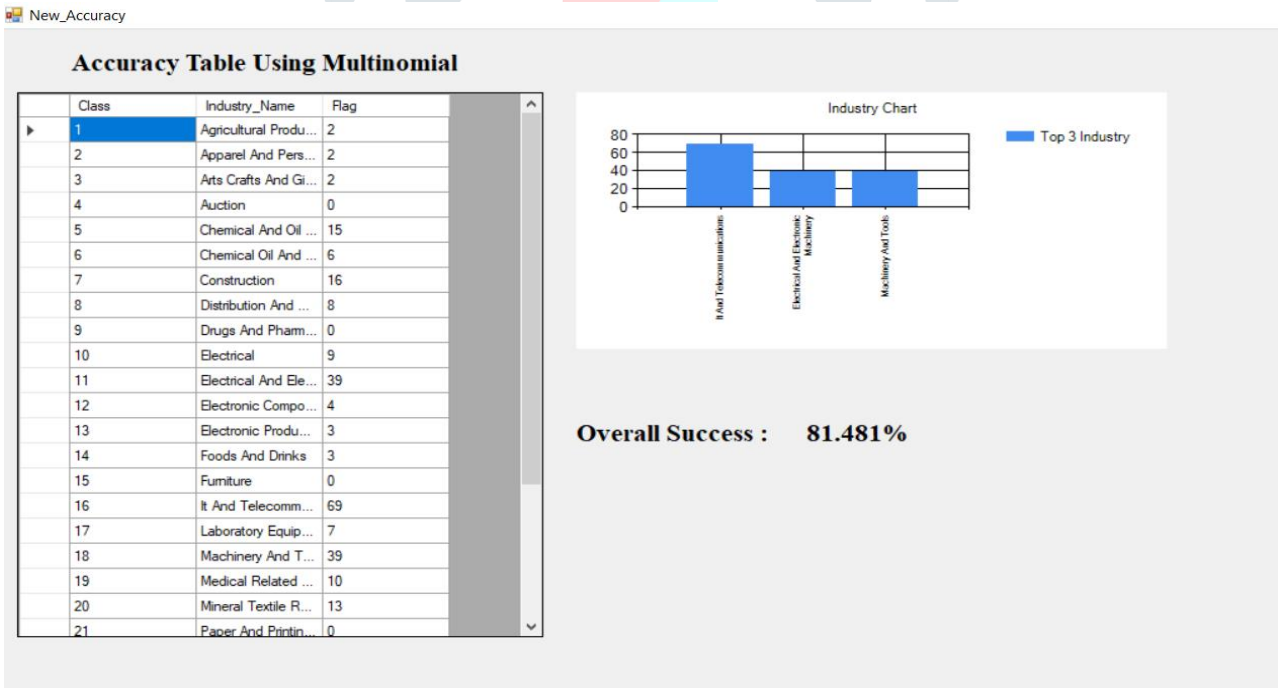

Fig-4.4: Proposed system output.

4.5 Confusion matrix for technological category classification

ConfusionMatrix

Confusion Matix For Industry Wise Classification

	Agricultural Product And Services	Apparel And Personal Care	Arts Crafts And Gift Article	Auction	Chemical And Oil Gas	Chemical Oil And Gas	Constr
▶	1	3	2	0	13	5	14
	2	0	0	0	15	7	8
	6	0	0	0	11	0	0
	2	0	0	0	3	0	4
	0	0	0	0	1	3	7
	0	0	0	0	0	0	1
	0	0	0	0	0	0	0
	0	1	0	0	0	0	0
	0	1	0	0	0	0	0
	1	0	0	0	0	0	0
	0	0	0	0	6	0	0

Fig-4.5: Confusion matrix for technological category classification.

4.6 Detail accuracy by class for technological category

Accuracy

Accuracy Table

	Class	TP Rate	FP Rate	Precision	Recall	FMeasure
▶	Agricultural Produ...	0.704	0.273	0.633	0.704	0.667
	Apparel And Pers...	0.852	0.789	0.548	0.852	0.667
	Arts Crafts And Gi...	0.926	0.913	0.521	0.926	0.667
	Auction	1.000	1.000	0.500	1.000	0.667
	Chemical And Oil ...	0.481	15.000	1.083	0.481	0.666
	Chemical Oil And ...	0.815	0.706	0.564	0.815	0.667
	Construction	0.704	0.273	0.633	0.704	0.667
	Distribution And ...	0.222	2.400	0.667	0.222	0.666
	Drugs And Pharm...	1.000	1.000	0.500	1.000	0.667
	Electrical	0.296	2.727	2.667	0.296	0.666
	Electrical And Ele...	0.074	2.087	0.095	0.074	0.670
	Electronic Compo...	0.481	15.000	1.083	0.481	0.666
	Electronic Produ...	0.296	2.727	2.667	0.296	0.666
	Foods And Drinks	0.148	2.211	0.267	0.148	0.664
	Furniture	1.000	1.000	0.500	1.000	0.667
	It And Telecomm...	0.296	2.727	2.667	0.296	0.666

Fig-4.6: Detail accuracy by class for technological category.

V. CONCLUSION

Here the proposed system shows that Selenium Automation can be used not only for testing purpose but it is used for Scrapping purpose also which will be helpful for many users for less time consuming and save manual efforts in browsing data from web pages. Similarly, by using Multinomial Naïve Bayes Classifier, this system can divide each and every document in an appropriate Technological category and improving the accuracy more from the existing methodologies.

REFERENCES

[1] https://en.wikipedia.org/wiki/Selenium_(software)
[2] https://stats.stackexchange.com/questions/33185/difference-between-naive-bayes-multinomial-naive-bayes
[3] https://www.geeksforgeeks.org/applying-multinomial-naive-bayes-to-nlp-problems/
[4] https://en.wikipedia.org/wiki/Bag-of-words_model.
[5] Sumit Goswami, Sunaina Kapoor, Prakriti Bhardwaj "Machine Learning for Automated Tender Classification" IEEE.
[6] Nicey Paul and Robin Tommy "An Approach of Automated Testing on Web Based Platform Using Machine Learning and Selenium" International Conference on Inventive Research in Computing Applications (ICIRCA 2018), IEEE 2018.
[7] Shakra Mehak, Rabia Zafar, Sharaz Aslam, Sohail Masood Bhatti "Exploiting Filtering approach with Web Scrapping for Smart Online Shopping" 2019 International Conference on Computing, Mathematics and Engineering Technologies – iCoMET 2019, IEEE 2019.
[8] Miroslav Bures, Martin Filipsky "SmartDriver: Extension of Selenium WebDriver to Create More Efficient Automated Tests", IEEE 2016.
[9] Satish Gojarea, Rahul Joshib, Dhanashree Gaigawarec "Analysis and Design of Selenium WebDriver Automation Testing Framework" 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), Procedia Computer Science ELSEVIER 2015.