

**CONTEXT-BASED UNDERSTANDING OF SCHOLARLY
ARTICLES AND STEM RESEARCH USING TEXT MINING
APPROACH**

By

**MD. ABDULLAH AL KAFI
ID: 191-15-12152**

**ISRAT JAHAN TASNOVA
ID:191-15-12932**

AND

**MD. WADUD ISLAM
ID: 191-15-12547**

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By
Dr. Sumit Kumar Banshal
Assistant Professor
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

Dhaka, Bangladesh

January 2023

APPROVAL

This Project is titled "Context-based understanding of scholarly articles and STEM research using text mining approach" Submitted by Md. Abdullah al Kafi, Id No:191-15-12152 ; Israt Jahan Tasnova, Id No: 191-15-12932 and Md. Wadud Islam, Id No: 191-15-12547 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 29.01.2023.

BOARD OF EXAMINERS



Chairman

Dr. Touhid Bhuiyan

Professor and Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Internal Examiner

Md. Abbas Ali Khan

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Internal Examiner

Ms. Aliza Ahmed Khan

Senior Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



External Examiner

Dr. Md. Sazzadur Rahman

Associate Professor

Institute of Information Technology

Jahangirnagar University

DECLARATION

We hereby declare that this project has been done by us under the supervision of Dr. Sumit Kumar Banshal, Assistant Professor, Department of CSE Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:



Dr. Sumit Kumar Banshal

Assistant professor
Department of CSE
Daffodil International University

Submitted by:



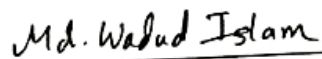
Md. Abdullah-Al-Kafi

ID: 191-15-12152
Department of CSE
Daffodil International University



Israt Jahan Tasnova

ID: 191-15-12932
Department of CSE
Daffodil International University



Md. Wadud Islam

ID: 191-15-12547
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing making us possible to complete the final year project/internship successfully.

We are really grateful and wish our profound indebtedness to Dr. Sumit Kumar Banshal, Assistant Professor, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Machine learning” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to Dr. Touhid Bhuiyan, Head, Department of CSE, and Ms. Subhenur Latif for their kind help to finish our project and to other faculty members and the staff of the CSE department of Daffodil International University.

We would like to thank our entire coursemates at Daffodil International University, who participated in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Publishers adopt different ways to sort the papers. Two of the most common tagging methods are Journal based tagging and Content-based tagging. In journal based tagging the tags are given to the papers depending on the published journal's interested fields. On the other hand, authors like Dimension use machine learning to classify and sort the paper's categories. Which method conveys more information and accuracy is a question to be answered. This study revealed that content base tagging is better than journal base tagging for sorting the paper's categorization by using title, abstract, and keywords. Gender discrimination or inequality refers to the unequal treatment of humans depending on their gender. To measure gender discrimination Gaye et al. used 3 statistical dimensions labor market, empowerment, and reproductive health. Education is one of the most important indicators of the gender discrimination index. Stereotype thinking plays a key role to demotivate females to participate in technical fields. World Bank data shows that the girl's participation ratio in primary and secondary education is increasing in developing countries like India, Bangladesh, Pakistan, Indonesia, and Nepal. On the other hand, gender bias has a negative impact on girls' education and choosing STEM-related subjects. Technical Education or a full form of STEM (STEM) related education plays a vital role in the development of a nation. This research aims to find out the veracity of stereotypical views and thinking. Besides, this research revealed female performance on technical and non-technical subjects and find that females are better than males in both technical and non-technical subjects.

TABLE OF CONTENTS

CONTENTS	PAGE
APPROVAL	
BOARD OF EXAMINERS	ERROR! BOOKMARK NOT DEFINED.
DECLARATION	ERROR! BOOKMARK NOT DEFINED.
ABSTRACT	V
CHAPTER	
CHAPTER : INTRODUCTION	1
1.1 INTRODUCTION	1
1.2 MOTIVATION	4
1.3 PROBLEM DEFINITION	4
1.4 RESEARCH QUESTIONS	5
CHAPTER 2: BACKGROUND	6
2.1 INTRODUCTION	6
2.2 RELATED WORKS	6
CHAPTER 3: RESEARCH METHODOLOGY	12
3.1 INTRODUCTION	12
3.2 DATA SET AND PREPROCESSING	12
3.3 MODEL INTRODUCTION	15
CHAPTER 4 : EXPERIMENTAL RESULT AND DISCUSSION	18
4.1 EXPERIMENTAL RESULTS	18
4.2 DISCUSSION	44

CHAPTER 5 : IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY	46
5.1 IMPACT ON SOCIETY	46
5.2 ETHICAL ASPECTS	46
5.3 SUSTAINABILITY PLAN	47
CHAPTER 6 : SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE	48
6.1 SUMMARY OF THE STUDY	48
6.2 CONCLUSIONS	48
6.3 LIMITATIONS	49
6.4 IMPLICATION FOR FURTHER STUDY	50
APPENDIX	51
ABBREVIATION:	51
APPENDICES: RESEARCH REFLECTION:	51
REFERENCE	52

LIST OF FIGURES

FIGURES	PAGE NO
FIGURE 1: PREPROCESSING STEPS.	14
FIGURE 2: INITIAL DATASET WITH 7 COLUMNS.	14
FIGURE 3: THE FINAL DATASET	14
FIGURE 4: DATASET DISTRIBUTION	15
FIGURE 5: WOS ABSTRACT	19
FIGURE 6: DIMENSION ABSTRACT	19
FIGURE 7 : WOS AUTHOR KEYWORD	20
FIGURE 8 : DIMENSION AUTHOR KEYWORD	20
FIGURE 9 : WOS ARTICLE TITLE	21
FIGURE 10 : DIMENSION ARTICLE TITLE	22
FIGURE 11: BASE MODEL VALIDATION ACCURACY ON DIMENSION DATASET.	23
FIGURE 12 : ACCURACY OF THE BASE MODEL ON THE WOS DATASET	24
FIGURE 13 : ACCURACY OF THE ENSEMBLE MODEL ON WOS AND DIMENSION DATASET.	25
FIGURE 14 : THE INTERFACE OF THE APP	27
TABLE 13 : YEARLY PARTICIPATION IN TECHNICAL AND NON-TECHNICAL SUBJECTS	29
FIGURE 15 : YEARLY PARTICIPATION RATIO OF FEMALE STUDENTS IN TECHNICAL SUBJECTS AND NON-TECHNICAL SUBJECTS.	30
FIGURE 16 : FEMALE PARTICIPATION IN TECHNICAL AND NON-TECHNICAL SUBJECTS CATEGORY 1	32
FIGURE 17 : FEMALE PARTICIPATION IN TECHNICAL AND NON-TECHNICAL SUBJECTS CATEGORY 2	34
FIGURE 18 : FEMALE PARTICIPATION IN TECHNICAL AND NON-TECHNICAL SUBJECTS CATEGORY 3	36
FIGURE 19 : MALE AND FEMALE PARTICIPATION RATIO IN TECHNICAL SUBJECTS FOR CATEGORY 1	38
FIGURE 20 : MALE AND FEMALE PARTICIPATION RATIO IN TECHNICAL SUBJECT FOR CATEGORY 2	40

FIGURE 21: MALE AND FEMALE PARTICIPATION RATIO IN TECHNICAL SUBJECT FOR CATEGORY 2	42
FIGURE 22: FEMALE CONTRIBUTION IN THE GOOD RESULT	44

LIST OF TABLES

TABLES	PAGE NO
TABLE 1 : NUMBER OF CLASSES (FIRST PART)	12
TABLE 2 : NUMBER OF CLASSES (SECOND PART)	12
TABLE 3 : TRADITIONAL MACHINE LEARNING DATA TRANSFORMATION METRICS	13
TABLE 4 : DEEP LEARNING DATA TRANSFORMATION METRICS	13
TABLE 5 : WOS VS DIMENSION ON ABSTRACT	18
TABLE 6: WOS VS DIMENSION AUTHOR KEYWORD	20
TABLE 7 : WOS VS DIMENSION ARTICLE TITLE	21
TABLE 8 : ACCURACY OF BASE MODELS ON DIMENSION DATASET	23
TABLE 9 : ACCURACY OF BASE MODELS ON THE WOS DATASET	24
TABLE 10 : ENSEMBLE MODEL ACCURACY	24
TABLE 11 : ENSEMBLE VS AVERAGE ACCURACY	25
TABLE 12. INTERFACE ELEMENTS	26
TABLE 13 : TOTAL PARTICIPATION	28
TABLE 15 : FEMALE PARTICIPATION IN TECHNICAL AND NON-TECHNICAL SUBJECTS CATEGORY 1	31
TABLE 16 : FEMALE PARTICIPATION IN TECHNICAL AND NON-TECHNICAL SUBJECTS CATEGORY	33
TABLE 17 : FEMALE PARTICIPATION IN TECHNICAL AND NON-TECHNICAL SUBJECTS CATEGORY 3	35
TABLE 18: MALE VS FEMALE PARTICIPATION RATIO FOR CATEGORY 1	37
TABLE 19: MALE VS FEMALE PARTICIPATION RATIO IN TECHNICAL SUBJECTS FOR CATEGORY 2	39
TABLE 20 : MALE AND FEMALE PARTICIPATION RATIO IN TECHNICAL SUBJECT FOR CATEGORY 3	41
TABLE 21 : FEMALE CONTRIBUTION IN THE GOOD RESULT	43

CHAPTER 1

INTRODUCTION

1.1 Introduction

The method used to sort the papers is called Tagging. Publishers adopt different ways to sort the papers. Two of the most common tagging methods are Journal based tagging and Content-based tagging. In journal based tagging the tags are given to the papers depending on the published journal's interested fields. On the other hand, in content-based tagging, the tags are given to the article depending on the content type. Publishers such as Web of Science use journal-based tagging to sort. Where the papers are sorted depending on the journal in which the paper is published. On the other hand, Publishers like Dimension use machine learning to classify and sort the paper's categories. Between the two methods which one conveys more information and accuracy is a question to be answered. The aim of this comparative study is to answer the question in a quantitative way.

An enormous number of Research articles, papers, and Journals are made available due to the recent development in Academic Research and Publications. Categorizing or sorting the papers according to their respective fields is hard. The method used to sort the papers is called Tagging. Publishers adopt different ways to sort the papers. Two of the most common tagging methods are Journal based tagging and Content-based tagging. In journal based tagging the tags are given to the papers depending on the published journal's interested fields. On the other hand, in content-based tagging, the tags are given to the article depending on the content type. Publishers such as Web of Science use journal-based tagging to sort. Where the papers are sorted depending on the journal in which the paper is published. On the other hand, Publishers like Dimension use machine learning to classify and sort the paper's categories. Between the two methods which one conveys more information and accuracy is a question to be answered. The aim of this comparative study is to answer the question in a quantitative way.

On the other hand, Gender discrimination or inequality refers to the unequal treatment of humans depending on their gender[1]. Regardless of significant advancement in recent years gender gap in privilege, prosperity, monetary and political sector continues in many developing and developed countries[2]. To measure gender discrimination Gaye et al. used 3 statistical dimensions labor market, empowerment, and reproductive health, where empowerment depends on educational attainment (Secondary and Tertiary education) and parliamentary representation [3]. On the other hand, parliamentary representation is heavily influenced by Higher Education [4] Education has a strong positive effect on Labor Market and women's reproductive health[5]–[12] So it is clear that education is one of the most important indicators of the gender discrimination index.

Female education has seen a lot of advancement in recent years though, women's participation is considered lesser and weaker and discrimination can be seen in the technical domains [7], [13]–[15]. Stereotype thinking plays a key role to demotivate females to participate in technical fields [16]. This report shows that females are underrepresented in the field of science and technology[17]. Technical fields such as Mathematics, Science, and technologies are fields of male dominance and a general perception has been noticed about more reliable performance from males in these different domains [13], [18]. The acronym STEM simply refers to Science, Technology, Engineering, and Math. Female students are less likely to enter Science, Engineering, Technical fields, or Mathematics[19]. Females are less likely to enter STEM-related jobs as well[20]–[22]. The reason behind this type of stereotypical thinking is imperfect information about women's performance in technical subjects[23].

Girls' education overview shows that the global primary and secondary education rates are nearly the same/almost on a similar scale for both males and females (male = 90% and female = 89%) [24]. Girls doing better than boys in high school is an international trend nowadays and more likely to get into university more than 70% of females and only 66% of males are getting into universities in Europe [21]. But the situation is different in developing countries where the income is low and roughly 36% of girls enroll in secondary education and approximately 11% in tertiary education[24]. World Bank data shows that

the girl's participation ratio in primary and secondary education is increasing for the developing countries like India, Bangladesh, Pakistan, Indonesia, Nepal[25], [26]. On the other hand, gender bias has a negative impact on girls' education and choosing STEM-related subjects[27].

On the other hand, In the Labour market people from STEM-related domains often got a higher pay scale than others. STEM-related professions are directly related to the economy[17], [21]. In tertiary-level education, STEM streams of studies are linked with a future job opportunities. So, a good understanding of the current situation of women in technical subjects is very much speculated. On the other hand, Technical Education or full form of STEM (STEM) related education plays a vital role in the development of a country. So, assessing female participation in STEM-related domains is taneed of the 21st century when almost all the sectors are eyeing to be in digital form.

So, the situation is Female education is on the rise and female students are doing better than their male counterparts, but male students are more likely to do better in technical subjects which raises questions such as, Is it true female students are doing badly in technical subjects or is it just stereotypical thinking? Or Are female students doing better in non-technical than technical subjects? Or Are female students in technical fields underrepresented?

According to a World Bank report, understanding the subject choice and enrollment of males and females in tertiary education is not enough to understand the situation, their performance difference in STEM-related subjects is also important [21]. To have a clear perspective on female participation in technical subjects, it is very much needed to have a comparative study of gender distribution and performance[21]. The comparative study will not only be able to give a perspective on the current situation but will also provide insight into the aspirations of the younger generation in the future market.

In this motivation, this study focuses to draw a line between the performance of females in the technical and non-technical subjects vis-a-vis male-female performance in technical subjects in terms of different indicators such as enrollment, participation, and contribution in performance academically.

1.2 Motivation

Our first objective was to obtain a research degree and the benefits that came with it. But in doing so, we ran across some challenging problems. We've noticed that there are several kinds of paper available online. However, very few of them are precise and to the point. Some author's works have been falsely tagged, or they don't correspond to the paper type. This slowed down our activity. Then, we sought to use a research tool to find a solution to this challenging problem. In order to benefit society as a whole. We first wanted to know which kind of labeling would function the best. Therefore, we discovered that content-based tagging performs better than journal-based tagging. Then, we tried to develop a tool that may assist the author in writing a paper that is relevant to their field of study. The research article domains may be verified using the method described in this study. In order to make their papers more closely connected to the targeted domain, writers may use this tool to examine and understand the target domains for their research paper and make any necessary modifications. With these instruments, we can contribute to society. This will assist us to establish our credibility.

1.3 Problem Definition

Machine learning and data mining are the two major terms in the research area. To give a proper solution it's necessary to find out the problems and related requirements in this field. It's also necessary to know the government policy or regulations and software industry requirements along with the course methodologies to implement machine learning in tagging. To determine the best categorization method between Journal based categorization and Content-based categorization we used Machine learning Algorithms. Going through a short survey of authors and researchers to find out the problems related to the paper's domain. In the current era, modern knowledge faces many significant changes in the achievement of tertiary education in developing the skills needed by the country. The participation of females in Higher Education especially in technical areas is increasing recently. The enrolment of female students in higher education can prove that female

students are getting occupied with the study field and doing better than male students. In effect, girls are outperforming boys in tertiary education. But segregation of gender or differences in enrolment in STEM courses still existed causing the enrolment difference in the technical education sector and STEM.

1.4 Research Questions

Here the main questions that this study will focus on are given below:

- Is content-based tagging retaining more correlation than Journal-based tagging?
- Is one method better than another with respect to data mining methodologies?
- Finding out the Real-life application and implementation of the study
- Authenticity of gender bias in the STEM and technical subjects about males and females in academia.

CHAPTER 2

BACKGROUND

2.1 Introduction

Due to modern technologies such as the Internet, the research and development field in different sections of science is experiencing huge improvement. This causes a lot of publications in different sections of science. So, in the era of automation, the process of indexing and sorting published articles and papers requires an automated solution. But no specific study has been conducted on finding the impact of automation. The method used to sort the papers is called Tagging. Publishers adopt different ways to sort the papers. Two of the most common tagging methods are Journal based tagging and Content-based tagging. In journal based tagging the tags are given to the papers depending on the published journal's interested fields. On the other hand, in content-based tagging, the tags are given to the article depending on the content type. Publishers such as Web of Science use journal-based tagging to sort. Where the papers are sorted depending on the journal in which the paper is published. On the other hand, Publishers like Dimensions use machine learning to classify and sort the paper's categories. Between the two methods which one conveys more information and accuracy is a question to be answered. The aim of this comparative study is to answer the question in a quantitative way.

2.2 Related Works

An enormous number of Research articles, papers, and Journals are made available due to the recent development in Academic Research and Publications. Categorizing or sorting the articles according to their respective fields is hard. Text classification also known as text tagging or text categorization is the process of categorizing text into organized groups.

Which can be helpful in the categorization of the Research articles. An increasing amount of research articles are being introduced regularly. New studies on research articles

categorization and classification are introduced as well[28]–[35]. [29] Used CNN, NBM, SVM, and KNN on research articles related to data science and data analytics from 2020 to 2022 of Scopus, ProQuest, and EBSCOhost depending on the proposed methods and content of the paper. In both cases, CNN outperformed other algorithms. A study was conducted on the library dataset to classify the library book using a bag of words on Random forest, SVM, and Decision tree. Random forest outperformed all the algorithms with an accuracy of 89%[30]. In another study of scholarly articles classification different machine learning and deep learning algorithms are applied [31]. They used tfidf with unigram and applied random forest, naïve Bayes, support vector machine, logistic regression, and CCNN, and DANN with a total number of classes 104. They targeted the abstract of the articles and the best performance was achieved by the CCNN. A binary classification study of science domain on 11778 articles from arxiv on three sub-domain of machine learning using bilstm, Asymmetric Word Embedding, and hierarchical attention networks and HAN outperformed other algorithms[36]. Bilstm and knowledge graph applied to scholarly article dataset to find out the relativeness and also shows how the combination of deep learning algorithms can improve performance while scholarly article classification[32]. Scope classification study of the scholarly article using abstract with 7 classes using BERT and ensemble learning algorithms with an f1 score of 91%[33]. On the other hand, an array of machine learning and deep learning algorithms are already used for text classification. Many studies [37] their studies applied multinomial and multivariant Naive Bayes algorithms in 5 different datasets from different sources and found that multinomial is 27%-50% better than multivariant. [38] proposed a transfer classifier based on the Naïve Bayes classifier applied to 3 different news datasets and found that transfer Naïve Bayes was better than SVM and traditional Naïve Bayes.[39] study of Naïve Bayes on 3 unbalanced news datasets showed that normalization significantly increases the accuracy [40] used naïve Bayes on 2 mail datasets with 45396 junk mail and 18314 normal mails. Depending on the number of features and auxiliary features, the accuracy increased by 85%-86% and 87%-88% as the number of features increased from 1000 to 2000.[41] their

comparison study of different Naïve Bayes algorithms such as Bayesian, Bernoulli, gaussian, and classical showed that the Bayesian method performed better in most of the datasets. [42] showed that the multinomial Naïve Bayes is better than Bernoulli on a news polarity dataset where the multinomial was 73.4% accurate and Bernoulli only 69.15%. [43] applied logistic regression, random forest, and k-nearest neighbors on the BBC news dataset with tfidf vectorizer, and logistic regression attains the highest accuracy of 97%. Another comparative study of traditional and deep learning algorithms on tobacco datasets conducted by [44] showed that logistic regression produced the best result among the traditional algorithms but the deep learning technique is better than traditional algorithms in all cases. [45] conducted Naïve Bayes and Logistic Regression on Twitter data and found that logistic regression performs better than Naïve Bayes. Logistic regression achieved 91% accuracy and Naive Bayes achieved 90% accuracy. [46] proposed a hybrid model based on CNN and SVM which outperformed both base algorithms [47] applied Naïve Bayes, SVM, and Logistic Regression on 3 different datasets, and found that SVM outperformed the rest of the algorithms. Ensemble classification methods are widely studied on different types of datasets such as image datasets and text datasets[48]–[54]. On the other hand, a lot of studies have been conducted on different aspects of scholarly articles. Studies such as [55]–[62] deals with the effect of gender in different aspects of academia and scholarly articles. The use of social media and relationships between social media and scholarly articles and its effects on paper citations are discussed in different studies[63]–[69]. Disciplinary variation of scholarly articles and altmetrics analysis are also studied [70]–[76].

On the other hand, There is a significant advancement against gender discrimination in recent years but a huge number of researchers found that there exists gender discrimination in different fields. The common fields of gender discrimination are privilege (access to basic rights), prosperity, the monetary and political sector in many developing and developed countries [77]. Education is one of the basic rights and considerable advancement in women’s education has been observed in recent times [13]. Though the

discrimination problem in gender is a well-observed issue in both developing and developed countries [78].

Stereotypical thinking in STEM-related to the thinking of males performs better than females in the STEM sectors [79]. Studying different STEM-related subjects in-depth shows some similar results. Thorson et al. in their paper show the effect of Stereotype thinking on math-related subjects and showed that there is a negative impact on women's performance when they are introduced to a stereotypical environment[80]. Williams et al. also talks about women underrepresentation in the field of science[81]. Stereotypical thinking in engineering is also being talked about in the paper where Singh et al. try to answer the about why women leave their jobs [82]. Schuster et al. showed in their paper that female contribution in science, technology, engineering and math (STEM) is less than male and they described the reason being the unfavourable stereotype of gender-biased thinking [16]. They proposed that a less stereotypical context regarding STEM is needed to solve this issue. A 2017 report from Global Gender Gap females are underrepresented in the field of science and technology[17]. This type of belief creates a false environment where discrimination spreads like wildfire and males are thought superior and good at science and technology whereas females are underrepresented despite legitimacy and talent [83]. Such stereotype belief affects an adverse effect while choosing a career[16]. Tests such as Draw a Scientist Test or DAST shows that Student from kindergarten to high school in the USA thinks of a scientist as a male person reference. Female enrolment in science and technology is increasing but still, stereotype thinking exists because of the underrepresentation of female characters[84]. Engineering is considered one of the most male-dominated sectors in the USA [85]. Engineering is considered a profession where the occupation of one sex is prominent. The first Engineering degree earned by a woman in the USA was back in 1892 but the current progress is very slow as the field still has a male dominance[82], [86], [87]. World Bank report shows that in Europe the situation is the same and these stereotypes also exist broadly [21], [88]. Nosek et al. Showed that the performance of students in math and science depends on an understanding of the topic not gender and male stereotype thinking is one of the main reasons for boys to choose STEM

carrier [89]. Stereotypical role model hinders females' anticipated success in STEM-related subjects [22]. Women exposed to stereotype thinking performed worse in Math and science [79]. So, stereotype thinking affects the performance of females in STEM.

Keng et al. identified imperfect information and stereotype thinking as one of the most important factors of gender bias/disparity [23]. It has been observed and discussed that, though the forced labour on women was decreasing the disparity still existed in Southeast Asian countries widely [1]. This type of phenomenon broadly exists in developing countries both in occupational and sectional segregation [90]. Socio-economic status also plays a vital actor in gender discrimination [2]. Education has always proven to have a positive effect on the socio-economic status of any individual [91]. Females in developing countries such as Pakistan, Indonesia, Bangladesh, India faces the most gender discrimination. Females in the education of most developing countries have faced a mass of challenges ranging from struggling to secure tuition, lack of moral and social support, minimum maintenance, and poor living conditions [92]. Naher et al. in their paper showed that the female students and teachers are underrepresented in different universities in Bangladesh depending on the enrolment and participation[93].

Paswan et al. studied the participation of women in higher studies and research. They showed that participation of women is found to vary in different disciplines, with biology (37%), agriculture science (32%), social science (31%) and medical science (32%) having a relatively higher number of female 1st authored papers as compared to engineering (20%), information science (21%) and mathematics (22%) [94]. This indicates that female students are less attracted to technical subjects. The most obviously striking feature of education in Pakistan is the prominence of gender inequality in education[95]. Shoaib et al. proposed the female students' outperformance and male students' underperformance in tertiary education at the University of the Punjab-Pakistan. In this paper, their Content analysis was carried out on the ten years results of master-level examinations. That is conducted from 2004-06 to 2013-15 [95]. The top three positions are secured by female students. Gabriel et al. focused on non-technical education. It shows female educational leaders and teachers at female dayah or Islamic boarding schools in Aceh. The study

explores the grassroots roles and motivation of female education leaders. This paper challenges societal inequalities and tensions[96]. Researchers tried to explore some findings from an Australian Faculty of Information and Communication Technology (FICT) against a backdrop of declining interest amongst women in courses and careers in Information and Communication Technology (ICT)[97]. They tried to theorise the educational dissociation where the pedagogical implications of these findings are consistently ignored in practice. Fokumet al. expressed that only 8% of females compared with over 50% of male undergraduates came to university ICT courses directly from high school[97]. World bank report shows that in most of the European and central Asian countries more females are participating in educational sectors[21].In primary and secondary education in developing countries of Asia female student participation is more than male student participation[25], [26].

These studies and reports heavily focus on only the participation ratio of females in the fields of technical studies. But it is important to measure the performance to understand the situation [21].

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This study is separated into 3 different domains. The first part is a ‘Descriptive Research’ of two types of data tagging methods namely content-based tagging and journal-based tagging. The second part is ‘Applied research’ where we apply the knowledge from the previous part and created a tool to classify research papers. The third or final part is ‘Exploratory Research’ where we tried to find the authenticity of gender bias in the STEM and technical subjects about males and females in academia.

3.2 Data Set and Preprocessing

For the First and second parts of the Study, the dataset was collected from the Web Of Science. experiments were conducted in 2 phases, each with 6 experiments on 3 types of data, namely the “Article Title, Abstract, and Author Keyword”. So, in each phase, 18 experiments are done on different input types with an 80-20 split where 80% of data is used for training and 20% for testing. The first phase deals with WOS categories and the second phase deals with Dimension Categories. In each phase, there were 2 types of algorithms applied. 1. Traditional machine Learning and 2. Deep learning.

Table 1 : Number of Classes (First part)

WOS	Dimension
61	64

Table 2 : Number of Classes (Second part)

WOS	Dimension
38	40

For the first part Naïve Bayes, Logistic Regression, Random Forest, Support Vector Machine

And for the second part Naïve Bayes, Logistic Regression, and Support Vector Machine from the traditional machine learning algorithms and 2 algorithms from deep learning namely Convolutional neural network and artificial neural network are used in both cases.

Traditional Machine learning: The Tfidf vectorizer was used with 5000 features and an n-gram range (1-3) for traditional machine learning algorithms.

Table 3 : Traditional Machine learning data transformation metrics

Vectorizer	Features	Stop word	n-gram Range
TFIDF	5000	English	1-3

Deep learning: For deep learning word embedding of 64 dimensions with 15000 frequent words was used.

Table 4 : Deep Learning data transformation metrics

Tokenizer	Embedding Dimension	Vocab size	Max Length	
TensorFlow Provided	64	15000	Abstract = 250	Other = 30

Each of the models is trained with the respective configuration mentioned in the previous section. The model's performance is collected and the models are retrieved for reuse as the base models for the Ensemble model. For traditional machine learning algorithms pickle[98] is used and for CNN and ANN h5 format is used.

The main two attributes of the dataset were the WOS categories namely Research Areas and Dimension Categories. Initially, the dataset consists of 76 columns, over 9 million data instances, and 178 Research areas. For these experiments, 7 columns were selected and only 1500 instances per class were selected after that the Dimension Categories were introduced to the instances depending on DOI. In the end, the number of instances was 114000 with 38 classes for WOS and 40 Classes for Dimension. While training 80-20 split was used for training and testing. Here Figure 1 shows several preprocessing steps.

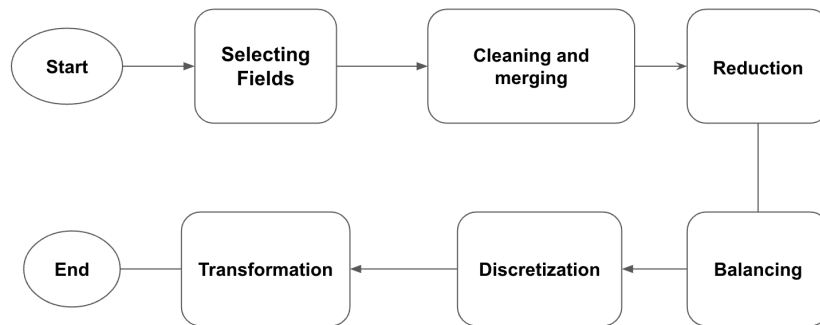


Figure 1: Preprocessing steps.

TI	DE	ID	AB	DI	WC	SC
A WSN-based automa	Honey bee; Colony; In	POLLINATION	The activity of a honey bee colony	10.1016/j.compag.2016.03.	['agriculture, multidisciplinary', 'computer science, interdisciplinary appli	Agriculture; Computer Science
Mapping interannual v	Normalized differenc	LAND-SURFACE PHENOLC	Accurate mapping interannual vari	10.1016/j.compag.2016.03.	['agriculture, multidisciplinary', 'computer science, interdisciplinary appli	Agriculture; Computer Science
Robust visual servo cor	Vision-based control; ORANGE PICKING	ROBOT	Unknown fruit motion due to exog	10.1016/j.compag.2016.03.	['agriculture, multidisciplinary', 'computer science, interdisciplinary appli	Agriculture; Computer Science
Calibration and validat	Calibration; APSIM-W NORTH CHINA PLAIN; CLIP	Crop growth in process based crop	10.1016/j.compag.2016.03.	['agriculture, multidisciplinary', 'computer science, interdisciplinary appli	Agriculture; Computer Science	
Determination of agric	Land consolidation (Li	GROUNDWATER LEVEL; S	Land consolidation (LC) is a technic	10.1016/j.compag.2016.03.	['agriculture, multidisciplinary', 'computer science, interdisciplinary appli	Agriculture; Computer Science

Figure 2: Initial dataset with 7 columns.

ArticleTitle	AuthorKeyword	KeywordsPlus	Abstract	DOI	WoSCategori	ResearchArea	Dimension
Minicharged p; solar physics; star	MAGNETIC DIPOL	We study the impact o	10.1088/1475-7516	['astronomy & Astronomy & As	0202	Atomic Molecular Nuclear Particle and Plasma Physics	
Mg line format line: formation; st	LATE-TYPE STARS	Context. Mg is the a e	10.1051/0004-6361	['astronomy & Astronomy & As	0202	Atomic Molecular Nuclear Particle and Plasma Physics	
Comparison of Spacecraft chargii	ENVIRONMENT; †	In the paper, we discuss	10.1016/j.asr.2015.	['astronomy & Astronomy & As	0202	Atomic Molecular Nuclear Particle and Plasma Physics	

Figure 3: The Final dataset

The third part of the study uses a dataset with 2 sections and 4 classes, and the total number of students is 8455 (Computer Science, Electrical, And Electronics Engineering). this study uses a tertiary level result data set of 8999 students over 5 years in 2 sectors (Technical and Non-Technical). Each of these sectors has two subjects/classes. In the technical sector, it has CSE And EEE and in the non-technical sector, it has BBA and LLB. the collected results are from the 1st semester to the 7th semester and the CGPA is calculated accordingly. and the dimension of the dataset is 8999*10. The columns represent id, sex, results from the 1st semester to the 7th semester, and CGPA.

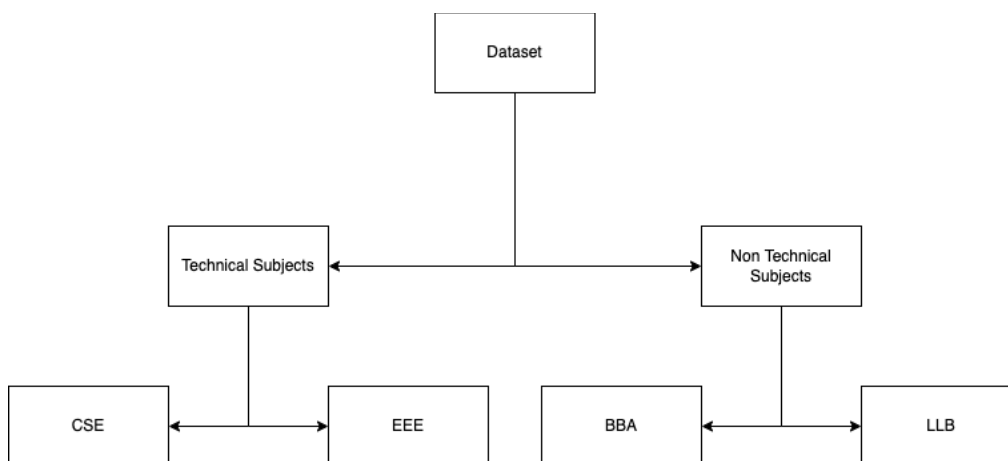


Figure 4: Dataset Distribution

It was necessary to pre-process the data because of dropouts and different sections in the dataset. The dataset was pre-processed in several steps. The dropouts were cleared from the dataset. Then the dataset was divided into 2 parts technical and non-technical. Then these 2 parts are divided into 4 different classes. After that, the female and male students from the datasets were extracted.

3.3 Model Introduction

Convolutional Neural Network

Convolutional Neural Network. A convolution neural network is an algorithm used to predict categorical cross entropy using a given set of independent variables. The 1D convolution model (Conv1D) has generated a convolution kernel that is convolved with the layer of input over one spatial aspect to the outcome of a tensor of outputs. The Conv1D model was configured with sigmoid as an activation function and the l2 regularizes for kernel regularization and the input padding were enabled. Here “Adam” is used as the optimizer with a learning rate of 0.005. The Word embedding layer provided by TensorFlow is used for CNN with a vocabulary size of 3000.x

Naïve Bayes

Naïve Bayes algorithm is an easy probability classifier. It’s evaluated a set of probabilities by calculating the frequency and sequence of values in a given dataset. The Naive Bayes

algorithm is a simple probability classifier. It calculates a set of probabilities by counting the frequency and combinations of values in a given data set. This classifier learns from training data. In this classifier, the conditional probability of each attribute A_i is given the class label C . Naive Bayes is applied to some data sets and the confusion matrix is generated for classes having possible values. For example in a news dataset the method follows :

$$\text{pr}[E/H] = \frac{N!}{n_1! n_2! \dots n_k!} \prod_{i=1}^k \frac{p_i^{n_i}}{n_i!}$$

Here $\text{pr}[E/H]$ is the probability of the document/news given its class H . and N is the number of words in the news. n_i is the time of occurrence of the word in the news. p_i is the probability of obtaining the word from the news concerning category H .

In this experiment Multinomial, Naïve Bayes is used.

Support Vector Machine

Support Vector Machine or SVM is a linear machine learning model for classification and regression models. Here, the support vector classifier is used for classification purposes with the ‘hinge’ loss from the machine learning library Sikitlearn[99]. The equivalent formula is:

$$\min w, b \frac{1}{2} w^T w + C \sum_{i=1}^n \max(0, 1 - y_i (w^T \phi(x_i) + b))$$

Logistic Regression

Logistic regression is an algorithm used to predict the dependent categorical variable using a given set of independent variables. The logistic regression model is extremely sensitive to “bad” data. “Bad” data pointing the outlying responses and extreme points in the design space(X). Logistic regression has been developed to fill the gap of constant variability. The model takes the natural logarithm of the odds as a regression function of the predictors. With 1 predictor, X , this takes the form $\ln[\text{odds}(Y=1)]=0+1X$. This equation shows that is

the model coefficient and describes the influence of predictor X on the logit. The fundamental equation of the generalized linear model is $g(E(y)) = \beta_0 + \beta_1 X$. It predicts the probability of the occurrence of an event by fitting data to a logit function.

Random Forest

A Random Forest classifier is an ensemble classifier that produces multiple decision trees. To decrease the correlation between decision trees, random forest considers controlling the term ρ^2 . ρ^2 is the main part of the variance.

$$I_j^2 = 1 - B_j = 1 - B_j^2(b)$$

where $I_j^2(b)$ is the relative importance of both decision trees. RF classifiers can successfully handle high data dimensionality and multicollinearity, being both fast and insensitive to overfitting. Paul et al proposed an improved RF that performs with minimum classification trees. RF is a type of machine learning called bootstrap aggregation or bagging. Combining results from multiple models is called aggregation (majority votes). By bagging Random forest algorithms gain better accuracy.

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Results

The aim of the first part is a comparative analysis of 2 types of tagging methods namely 'journal-based tagging' and 'content-based tagging'. For this purpose, Naïve Bayes, Logistic Regression, Random Forest, Support Vector Machine from the traditional machine learning algorithms, and 2 algorithms from deep learning namely Convolutional neural network and artificial neural network are selected. 6 different experiments were done on the datasets to find out the performance of the algorithms in the dataset. For this Abstract, the Article title and Author Keyword are selected from the dataset and 2 types of tagging namely WOS tags and dimension tags are selected.

Comparing the accuracy from the Abstract it can be seen that content-based tagging shows better accuracy in each algorithm. The accuracies are shown in Table 5 and Figures 5, and 6.

Table 5 : WOS vs Dimension on Abstract

Model	WOS(x1)	Dimension(x2)	Difference (x2-x1)
Naïve Bayes	50.0	61.76	11.76
Logistic Regression	63.0	70.66	7.66
Random Forest	69.0	70.99	1.99
SVM	66.0	72.10	6.10
CNN	63.0	66.90	3.90
ANN	59.0	63.34	4.34

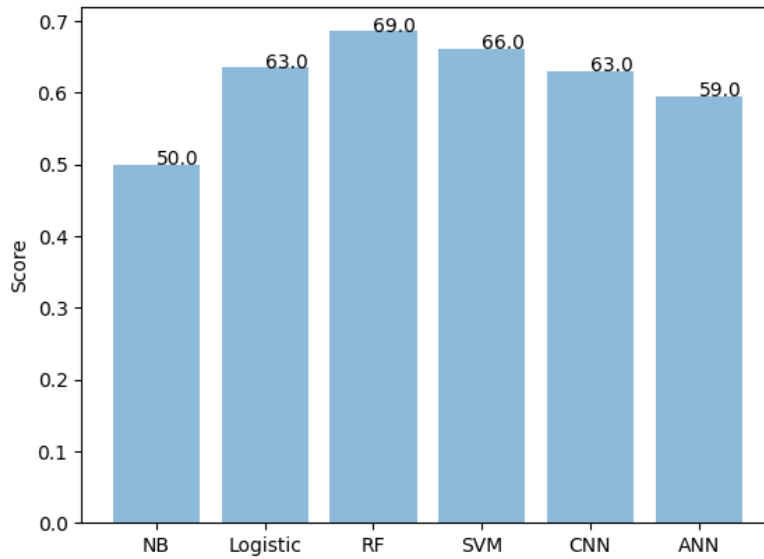


Figure 5: WOS Abstract

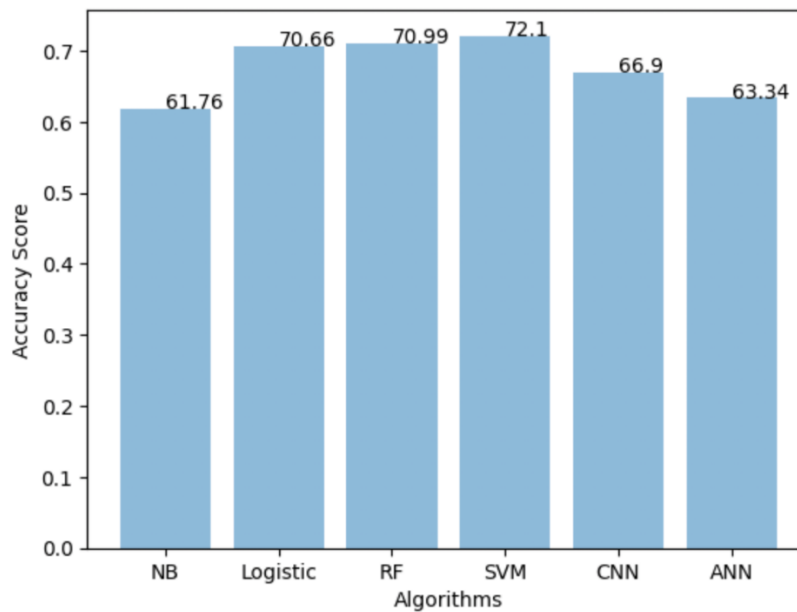


Figure 6: Dimension Abstract

From the table and the figures, it can be seen that the algorithms performed better on the content-based tags.

The next experiment deals with the Author keyword. The accuracy results are depicted in figures 7, 8, and table 6.

Table 6: WOS vs Dimension Author Keyword

Model	WOS(x1)	Dimension(x2)	Difference (x2-x1)
Naïve Bayes	50.00	53.41	3.41
Logistic Regression	53.81	57.40	3.59
Random Forest	62.04	63.94	1.90
SVM	55.85	58.93	3.08
CNN	56.19	57.54	1.35
ANN	54.01	56.66	2.65

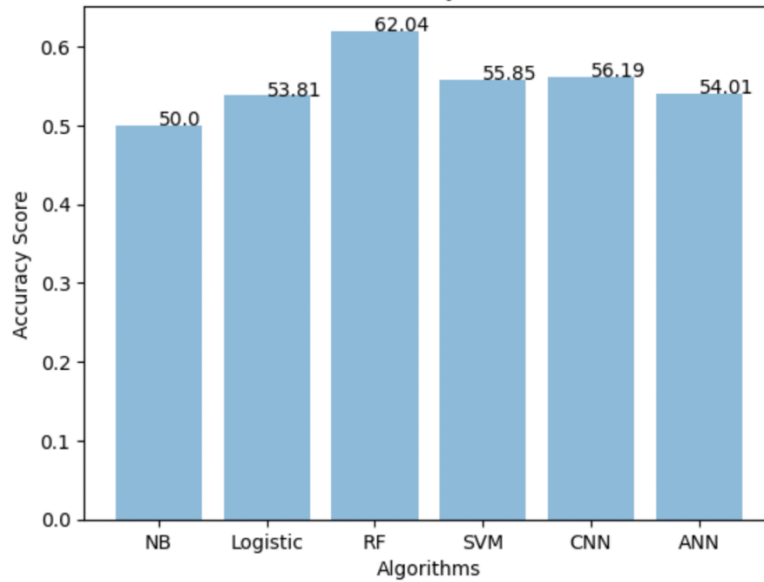


Figure 7 : WOS Author Keyword

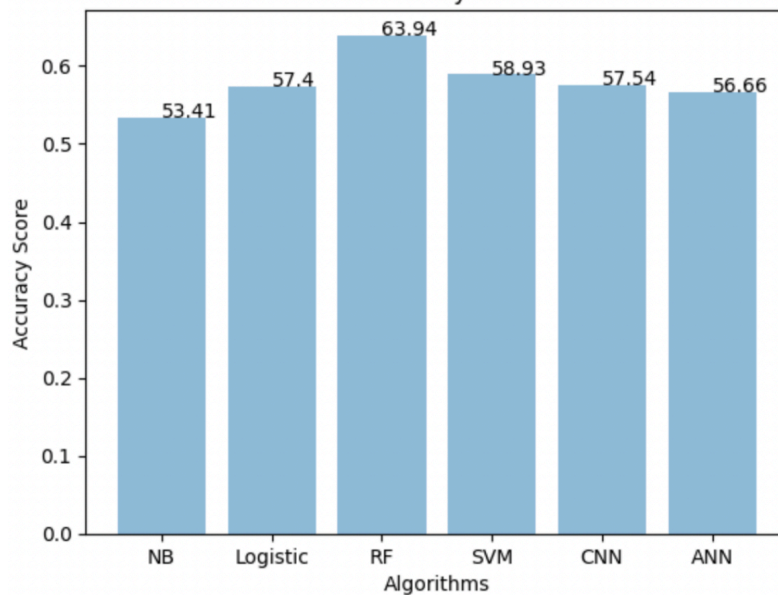


Figure 8 : Dimension Author Keyword

From the figure and the table, it can be seen that the algorithms performed better on the dimension tags again.

The next experiment deals with the accuracy of the models in the article title. Table 7 and figure 9, 10 depicted the results.

Table7 : WOS Vs Dimension Article Title

Model	WOS(x1)	Dimension(x2)	Difference (x2-x1)
Naïve Bayes	50.00	52.13	2.13
Logistic Regression	52.00	57.61	5.61
Random Forest	59.00	63.29	4.29
SVM	54.00	59.39	5.39
CNN	54.00	58.09	4.09
ANN	51.00	56.25	5.25

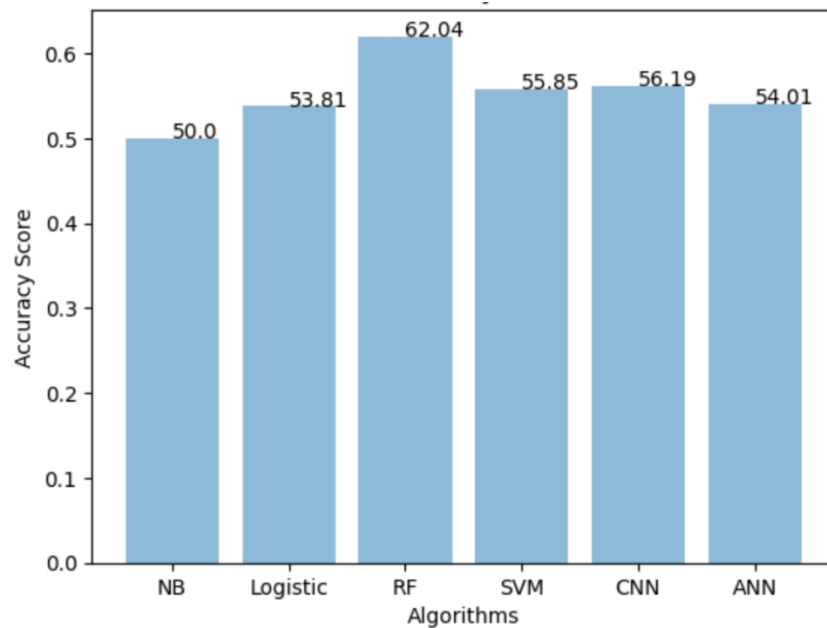


Figure 9 : WOS Article title

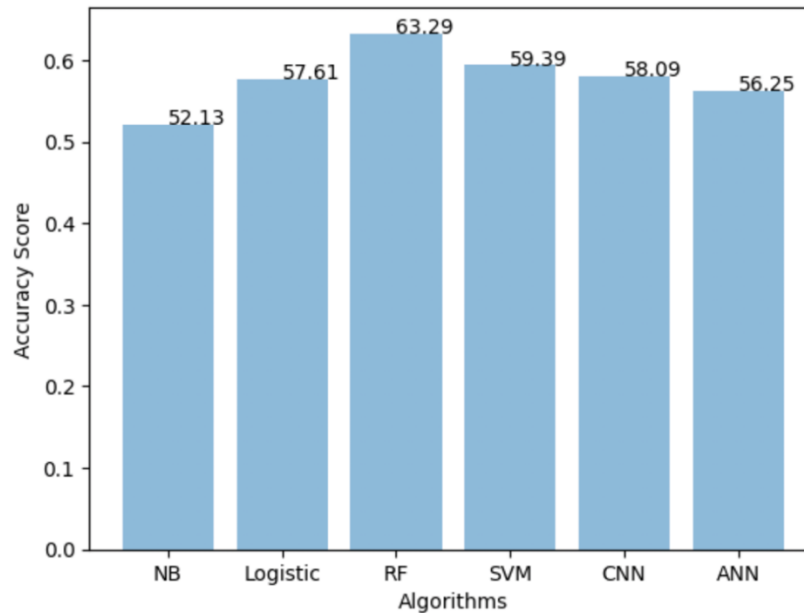


Figure 10 : Dimension Article Title

Here, Again the algorithms performed better on the dimension dataset. So, it can be seen that all the experiment setups' content-based tagging outperformed the journal-based tagging. It can also be seen that the content-based tagging retains more correlation and more useful to find and sort article.

Second Part

In the second part, as discussed in the previous section this study of the “Classification and Recommendation system” uses NB, LR, SVM, Conv Net, and ANN were selected. The experiments are conducted on Article Title, Author Keyword, and Abstract. 10 different experiments were done on the datasets to prepare the base models and 3 more experiments to measure the aggregated ensemble model. For better interpretation, the experiments will be clustered/packed in sets. Where the first set of experiments will deal with the base model’s performance on Dimension data, the second set of experiments will deal with the base model's performance on WOS data. The Third set of experiments will show the accuracy difference of the Ensemble model depending on the input type. From this point, the experiment sets will be addressed as exp1, exp2, and exp3 respectively. In each case,

only the validation accuracy is compared and experiment configurations and conditions are kept the same for each experiment.

Table 8 : Accuracy of base models on Dimension Dataset

Model	Abstract (%)	Author Keyword (%)	Article Title (%)
Naïve Bayes	76.84	67.03	65.80
Logistic Regression	80.46	67.25	66.23
Support Vector Machine	79.68	66.03	64.63
CNN	76.35	66.00	66.02
ANN	75.35	65.64	64.55

Here exp1's result has been represented in terms of validation accuracy on the Dimension dataset depending on all 3 input types i.e., 'Abstract', 'Keyword', and 'Article Title'. It is visualized in Table 8 and Figure 11 with their respective Accuracy as Scores (in percentage).

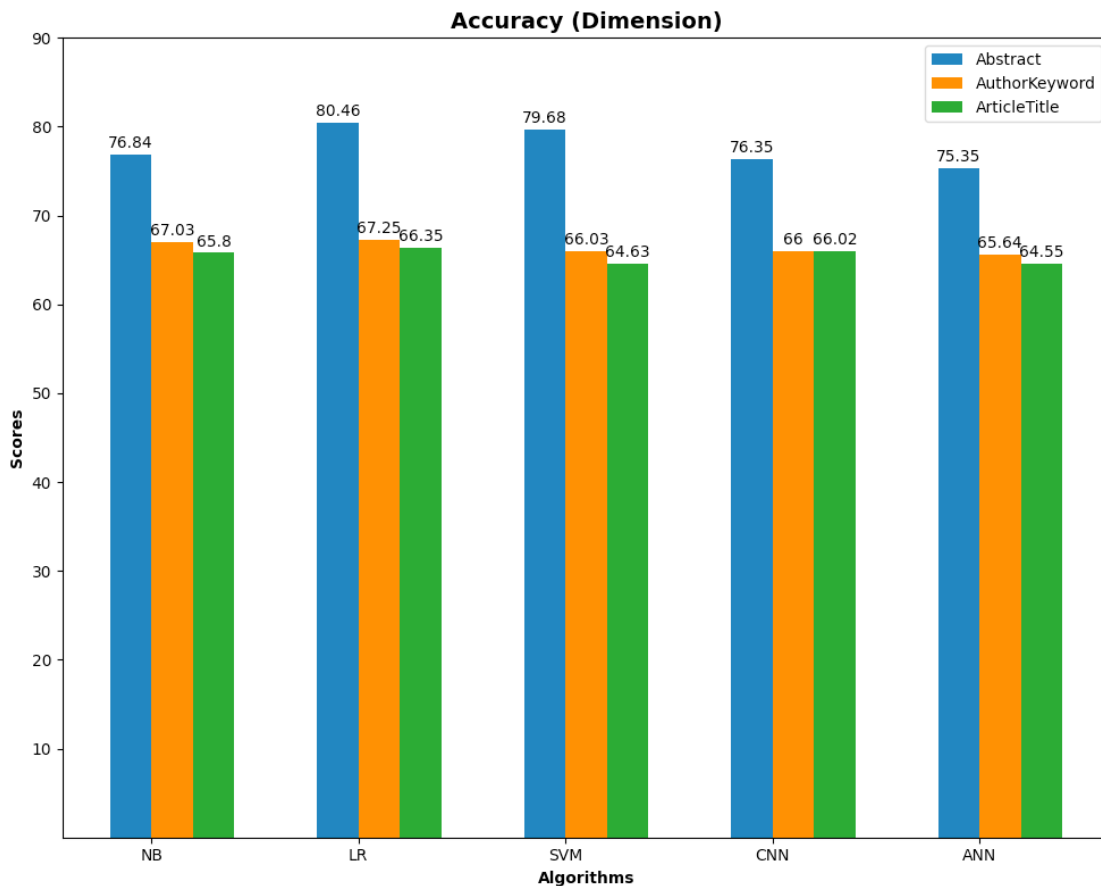


Figure 11: Base model validation accuracy on Dimension Dataset.

Table 9 : Accuracy of base models on the WOS Dataset

Model	Abstract (%)	Author Keyword (%)	Article Title (%)
Naïve Bayes	71.14	60.78	64.07
Logistic Regression	76.53	64.81	66.83
Support Vector Machine	76.61	64.96	67.27
CNN	75.25	64.99	67.76
ANN	74.34	63.72	66.07

In continuation of the experiments, exp2 carries all the attributes of exp1 but on a separate dataset namely the WOS Dataset. Here the result is visualized in Table 9 and Figure 12 with their respective Accuracy as Scores (in percentage).

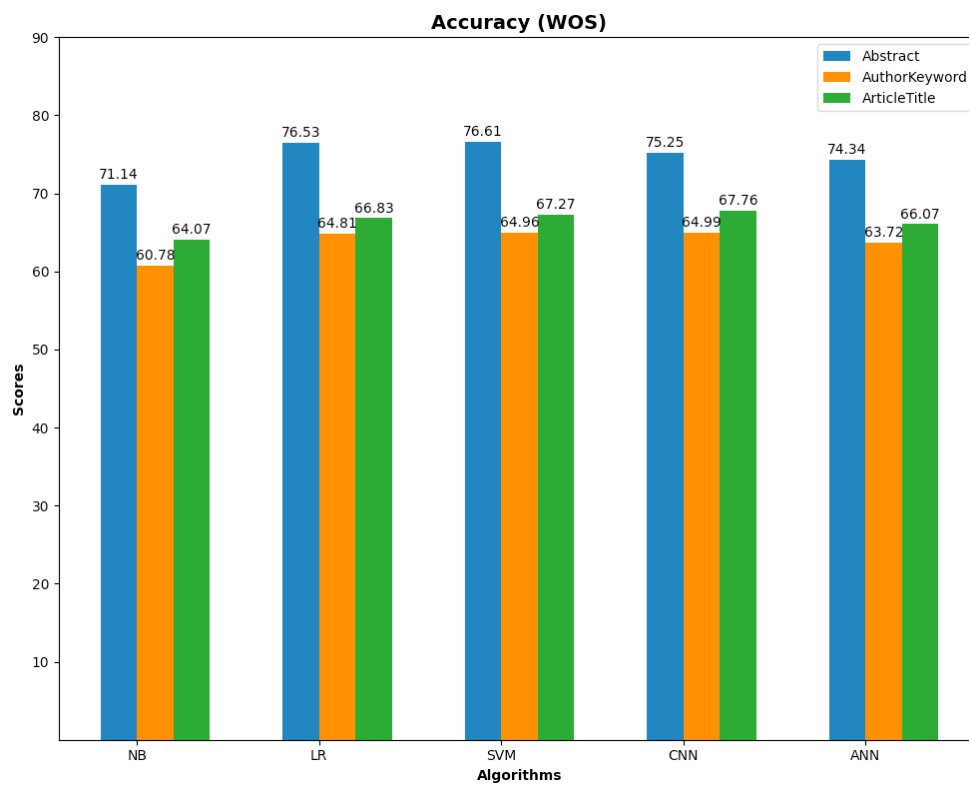


Figure 12 : Accuracy of the base model on the WOS dataset

The exp3 Deals with the proposed ensemble model and its performance in both datasets i.e., WOS and Dimension.

Table 10 : Ensemble model accuracy

Type	Abstract	Author Keyword	Article Title
WOS	79.10	68.20	70.10
Dimension	84.20	72.50	69.50

Here the result is visualized in Table 10 and Figure 13 with their respective Accuracy as Scores (in percentage).

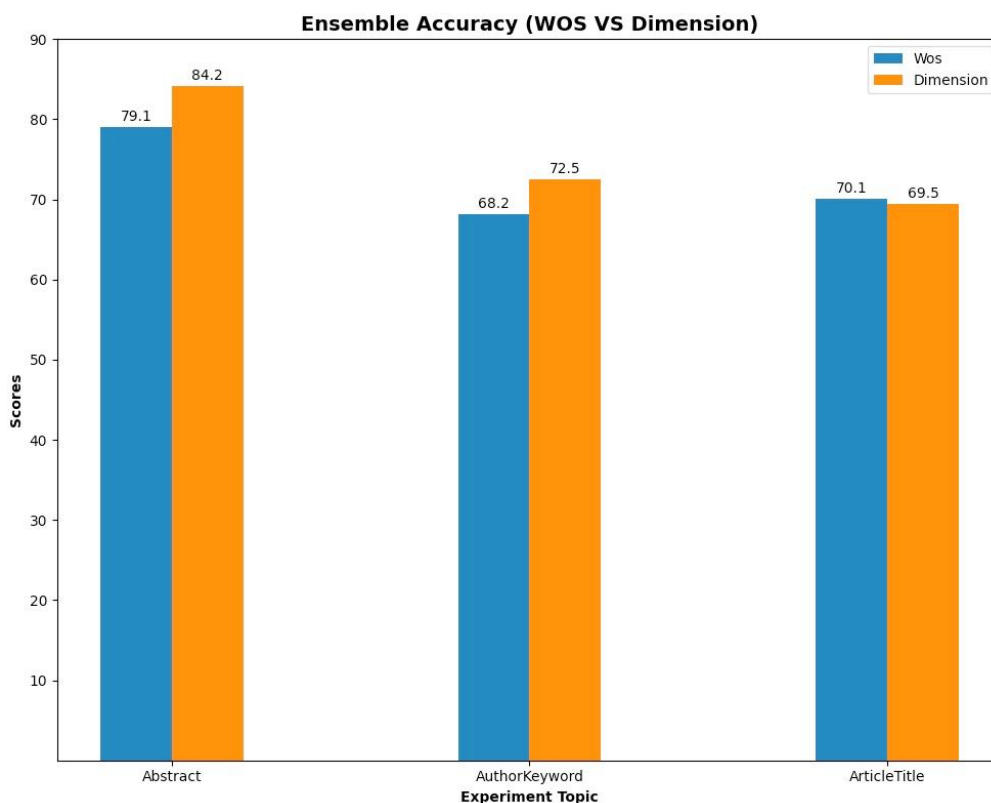


Figure 13 : Accuracy of the Ensemble model on WOS and Dimension Dataset.

Comparing base model performances to ensemble model performance shows a significant improvement in each input type.

Table 11 : Ensemble Vs Average Accuracy

Model	Abstract	Author Keyword	Article Title
Average Dimension(X1)	77.44	66.37	65.47
Ensemble Dimension(X2)	84.20	72.50	69.50
Difference (X2-X1)	6.76	6.13	4.03
Average WOS(X1)	74.77	63.85	66.40
Ensemble WOS(X2)	79.10	68.20	70.10
Difference (X2-X1)	4.33	4.35	3.7

Table 11 shows the difference between ensemble model accuracy and the average accuracy of the base models in each dataset. It can be seen that while experimenting with abstract and author keywords the ensemble algorithm achieves more than 6% and 4% accuracy in both Dimension and WOS datasets respectively. The increase in accuracy on article titles

is not as significant as the previous 2 but still, the improvement is more than 4% and 3% in both datasets.

The Web Interface

For the real-life applicability of the proposed model, the model is implemented in a web interface. The total web interface is a single-page interface so that the user can use it easily. The web interface can be used to see the correlation between the User's Article Title, Author Keyword, and Abstract with the predicted category and also get recommendations about the publishers who publish the same category articles. Figure 14 shows the web interface which has 8 elements.

Interface Elements	Function
Enter Your Title	This option takes the title of the user's article
Enter Keyword	This option takes the keywords selected by the users.
Enter your Abstract	This option takes the Abstract of the user's paper.
Radio Button	This button allows the user to choose the type of model he or she wants to use. Namely WOS or Dimension.
Sidebar	The sidebar shows the available classes for the selected model type.
Compute Button	It runs the algorithm and shows the prediction table and recommendation list.
Prediction table	The prediction table shows the predicted class and the confidence of the algorithm. Confidence values can be used to determine the correlation.
Common Publishing sources	This list recommends the user similar publishing sources for the predicted category.

Table 12. Interface Elements

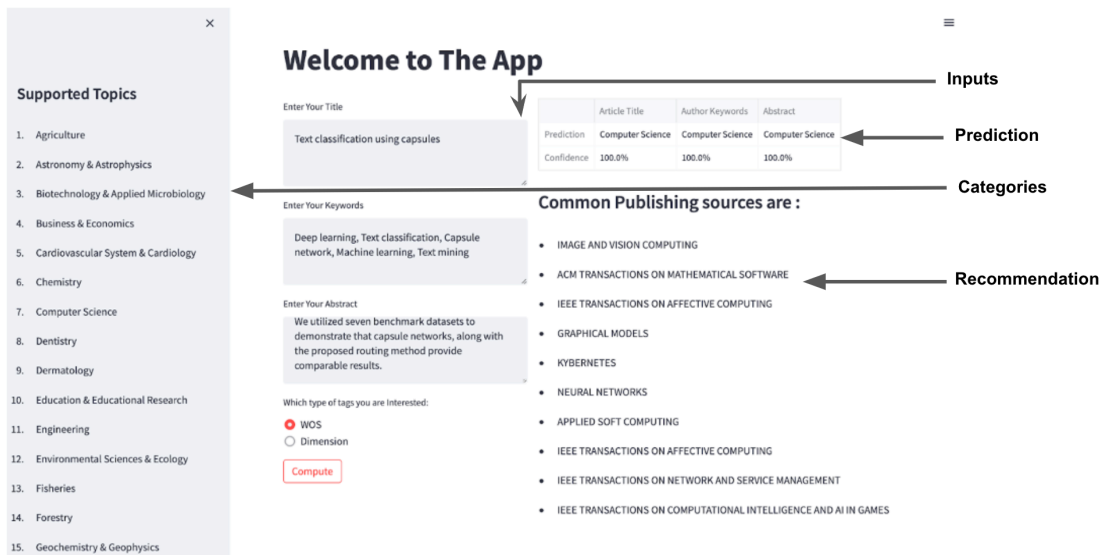


Figure 14 : The interface of the app

Third Part

In the third part, this study deals with investigates the current state of female students in technical subjects with respect to non-technical subjects. This paper mainly focuses on the enrolment status, results, and performance of female students and male students in different categories.

Here, the results are subdivided into 3 categories and category 1 is considered as a good result, category 2 is considered as an average result and category 3 is considered as a below-average result. The students are divided into 3 groups, Group 1 is female in technical subjects, Group 2 is female in Female in Non-Technical fields, Group 3 is Male in Technical fields.

Depending on Participation:

Number difference of Female students in Technical and Non-technical Subjects as Experiment 1

Depending on Student Result:

Female vs Female in Technical vs Non-technical Subjects as Experiment 2

Category 1: Result ≥ 3.5

Category 2: Result ≥ 3.0

Category 3: Result ≥ 2.5

Male vs Female in Technical Subjects as Experiment 3

Category 1: Result ≥ 3.5

Category 2: Result ≥ 3.0

Category 3: Result ≥ 2.5

Female Contribution in Total Result in Technical Subjects as Experiment 4

Experiment 1

The total number of students from the year 2015-2019 is 8999 of the 8455 are regular and 544 are irregular or dropped semesters. So, they were discarded in data pre-processing. Out of 8455 students, 7077 are from technical subjects and 1378 are from non-technical subjects.

Total female participation in technical subjects is 1262 which is only 17% of total participation. And in Non-technical subjects' participation is 459 which is 33.30%.

Table 13 : Total participation

Subjects	Total	Female	Percent
Technical	7077	1262	17%
Non-Technical	1378	459	33.3%

So, female participation in the non-technical subjects is higher than in technical subjects. And the participation ratio is approximately double in non-technical subjects. To have a clear view of the change in participation yearly participation ratio is important. And if we look at the yearly participation ratio the scenario changes a lot.

Table 14 : Yearly participation in Technical and Non-technical Subjects

Year	Group 1 (x1)	Group 2 (x2)	Difference (x2-x1)
2015	0.14	0.27	0.13
2016	0.15	0.34	0.19
2017	0.19	0.26	0.07
2018	0.22	0.43	0.21
2019	0.23	0.41	0.18
AVG	0.19	0.34	0.16
S.D	0.040	0.078	0.056

Regarding Female participation in technical subjects (Table 13; Group 1 (x1)), the standard deviation is 0.040 and the average is 0.19. From the calculation of average, it makes known that the yearly ratio of data is less far away with respect to average as the standard deviation is 0.040. The yearly ratio data are close to each data [100]. It points out that every year female participation in technical subjects is almost the same. Not decreased much in any year. This is a good sign because in the future female participation in technical subjects can increase.

In terms of Female participation in non-technical subjects (Table 13; Group 2 (x2)), the standard deviation is 0.078 and the average is 0.34. Group 2 (x2) participation ratios are widely spread out with respect to their average. By this calculation, it is displayed that every year female participation in non-technical subjects is varying widely. Therefore, female participation in non-technical subjects has no stability in increasing the participation rate. As a result, Group 1(x1) has the stability to enlarge the rate of participation more than Group 2(x2).

Female in Technical Vs Non Technical Subjects Participation

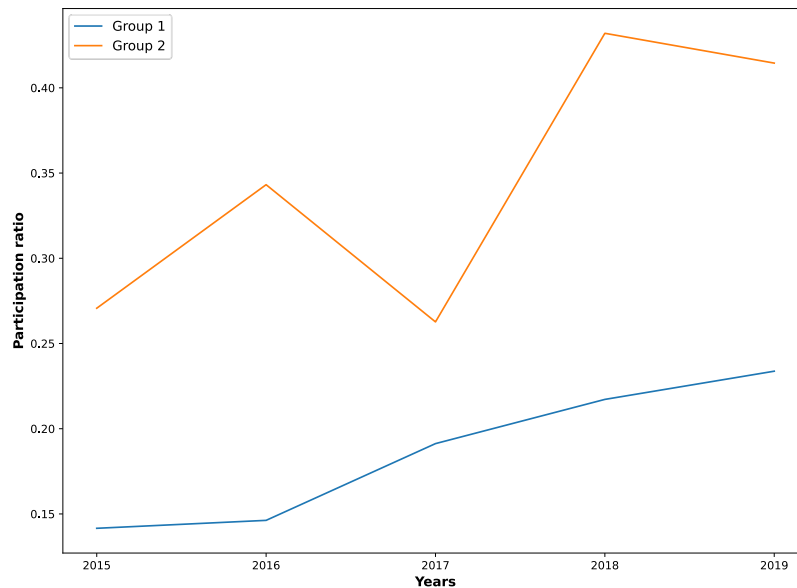


Figure 15 : Yearly participation ratio of Female students in Technical Subjects and Non-technical subjects.

From Table 13 and Figure 15, it can be seen that female participation in technical subjects is increasing. Although the participation ratio is lower than group 2. but each year the ratio increased considerably. The difference column from table 13 shows that the average difference is 0.156. So, on average, only 15% more females enrolled in the non-technical subjects. And the tables also show that the lowest difference is in 2017 when the enrolment in group 2 decreased to only 0.26. It was 30.76% lower than the previous year. but the enrolment in group 1 increased by 21.05%. In 2018 group 2 reached its peak participation ratio and the highest difference can be seen there. From Figure 15, it is clear that group 1 has a stable increase in the participation rate.

So, it is clear that the enrolment in group 2 is larger than the enrolment in group 1. The incremental yearly enrolment ratio in Group 1 is stable but group 2 shows a lot of variances.

Depending on Result

In this section, the aim is to check the participation ratio of female students according to the results to find out their current situation.

Experiment 2

In this experiment the aim is to find the participation of Female students Technical and Non-technical Subjects depending on results in 3 different categories.

Category 1: It focuses on the participation ratio of students with a result higher or equal to 3.50

Table 15 : Female Participation in Technical and Non-technical Subjects Category 1

Year	Group 1 (x1)	Group 2 (x2)	Difference (x2-x1)
2015	0.23	0.25	0.02
2016	0.23	0.31	0.08
2017	0.29	0.35	0.06
2018	0.33	0.30	-0.03
2019	0.43	0.23	-0.20
Average	0.3	0.29	-0.01
S.D	0.083	0.048	0.112

In the portion of female students with a result higher or equal to 3.50 in technical subjects (Table 14 ; Group 1 (x1)), the standard deviation is 0.083 and the average is 0.3. It expands the yearly ratio of data are marginally spread out with respect to average as the standard deviation is 0.083. The yearly ratio data are close to each ratio data [100]. It points out that every year the participation ratio of female students with a result higher or equal to 3.50 in technical subjects is nearly alike. The S.D indicates that in the future female participation in technical subjects can increase.

The proportion of female participation in non-technical subjects with a result higher or equal to 3.50(table 14 ; Group 2 (x2)), 0.29 is the average and the standard deviation is 0.048. Group 2 (x2) ratio data are extensively spread out with respect to its average. The calculation present, every year female participation in non-technical subjects with a result higher or equal to 3.50 is fluctuating broadly. That being so, female participation in non-

technical subjects with a result higher or equal to 3.50 has no solidity in increasing the participation rate. As a result, Group 1(x1) has secureness for the increasing rate of participation on technical subjects more than the participation in non-technical subjects with a result higher or equal to 3.50. Consequently, females are doing better in technical subjects

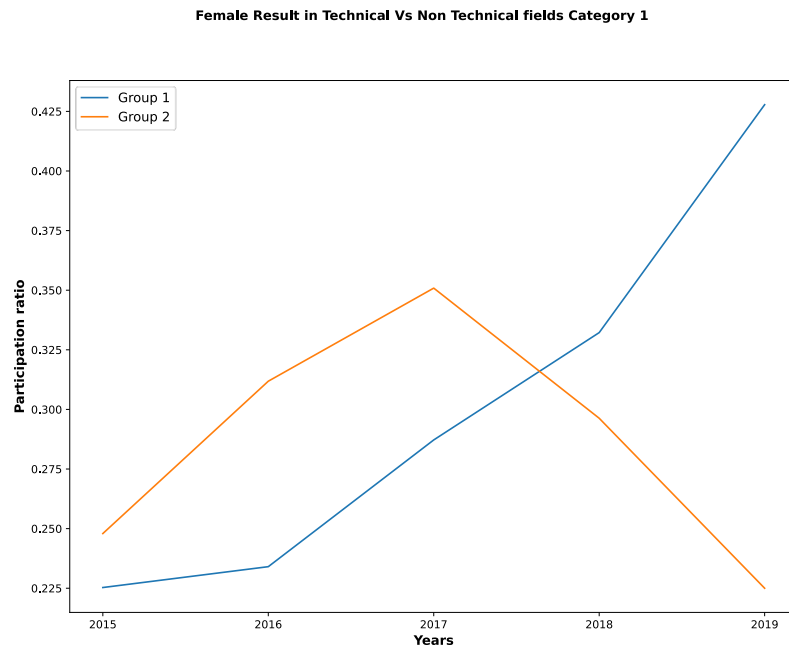


Figure 16 : Female participation in Technical and Non-technical Subjects Category 1

From Table 14 and Figure 16, it is clear that the number of females in technical subjects is doing better and is increasing each year from 2015 to 2019 as the performance curve is upward sloping. On the other hand, the female results in non-technical fields increased till 2017 then started to decrease. The difference column of the table shows the same. The rate of change in results in technical subjects is very high from 2017 to 2019 as the rise of the resulting curve in the figure is very high.

Considering the previous Table 15 and Figure 17 it is clear that from 2017 to 2019 the enrolment ratio in group 2 increased but here the resulting ratio decreased. On the other hand, the participation ratio in group 1 increased, and the result was as well. So it is clear

that although the participation ratio of group 1 is lower than group 2 the resulting ratio is higher and both participation and result ratio is increasing.

So, females in technical subjects are continuously doing better than females in non-technical subjects.

Category 2: It focuses on the participation ratio of students with results less than 3.50 and greater than or equal to 3.00 is present.

Table 16 : Female Participation in Technical and Non-technical Subjects Category

Year	Group 1 (x1)	Group 2 (x2)	Difference (x2-x1)
2015	0.34	0.4	0.06
2016	0.38	0.33	-0.05
2017	0.36	0.3	-0.06
2018	0.37	0.41	0.04
2019	0.44	0.24	-0.2
Average	0.37	0.34	-0.04
S.D	0.038	0.070	0.103

The number of female participation increased 2015-2019(Table 15 ; Group 1 (x1)). In the terms of female participation in technical subjects, the standard deviation is 0.038 and the average is 0.38. This revealed the yearly ratio of data is widely spread out with respect to average as the standard deviation is 0.038. The yearly ratio data are far from each ratio. It points out that every year the participation ratio of female students with a result less than 3.50 and greater than or equal to 3.00 in technical subjects is not alike. The female participation ratio in technical subjects has upright consistency in Table 15 , Group 1 (x1). With regards to (Table 16 ; Group 2 (x2)), the average is 0.34 and the standard deviation is 0.070. In Group 2 (x2) the difference between s.d and average is less. It indicates that here in Group 2 (x2), female participation on non-technical subjects are more than in Group 1 (x1). Still Group 2 (x2) participation ratios have less consistency. In order that, female participation in technical subjects with a result less than 3.50 and greater than or equal to

3.00 have quality participation rate according to ratios consistency. As a result, Group 1(x1) has secureness for an increasing rate of participation in technical subjects more than the participation in non-technical subjects with a result less than 3.50 and greater than or equal to 3.00.

Result Female in Technical Vs Non Technical fields Category 2

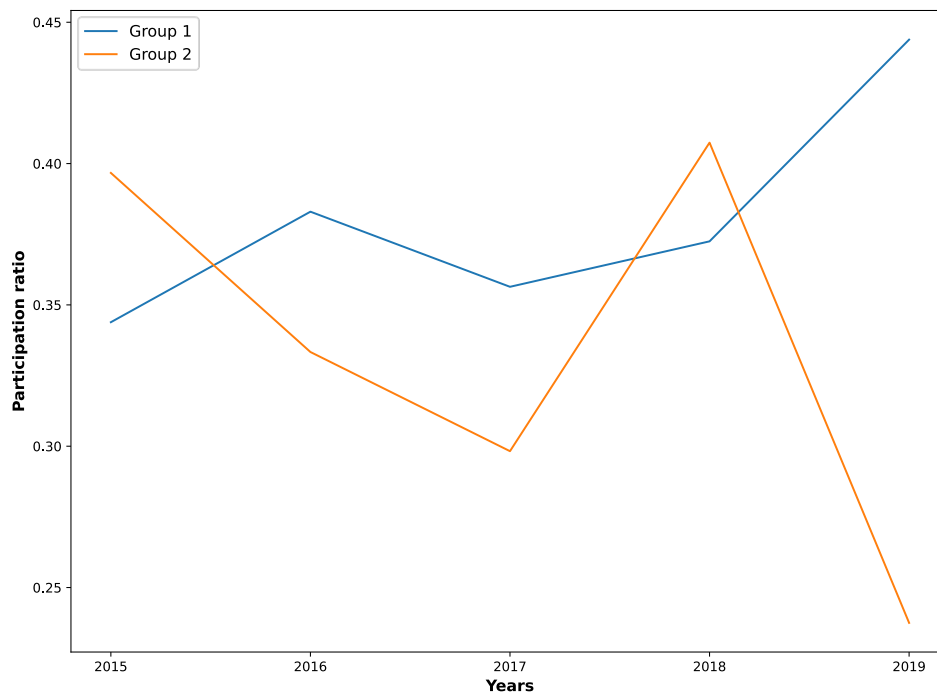


Figure 17 : Female participation in Technical and Non-technical Subjects Category 2

From Table 15 and Figure 17 , it can be seen that group 2 started well in 2015 and the ratio difference is 0.06. But after that, the ratio started to fall and only saw a rise in 2018 and the average difference is -0.21 which means that on average group 1 did better than group 2. In 2019 the difference is maximum. The figure shows that in 2019 the ‘Group 2’ line is downward facing and the ‘Group 1’ line is upward-facing and the distance between is noticeable. So, in category 2 group 1 is doing better than group 2 as well.

Category 3: It focuses on the participation ratio of students with results less than 3.00 and greater than or equal to 2.50.

Table 17 : Female Participation in Technical and Non-technical Subjects Category 3

Year	Technical (x1)	Non-Technical(x2)	Difference(x2-x1)
2015	0.25	0.27	0.02
2016	0.28	0.27	-0.01
2017	0.25	0.25	0.00
2018	0.20	0.22	0.02
2019	0.11	0.36	0.25
Average	0.22	0.27	0.06
S.D	0.066	0.052	0.109

In the portion of female students with a result less than 3.00 and greater than or equal to 2.50 in technical subjects (Table 17 ; Group 1 (x1)), the standard deviation is 0.066 and the average is 0.22. It disclosed that the yearly ratio of data is marginally spread out with respect to average as the standard deviation is 0.066. The yearly ratio data are close to each data [100]. It points out that every year the participation ratio of female students with a result less than 3.00 and greater than or equal to 2.50 in technical subjects is nearly alike. The S.D indicates that in the future female participation in technical subjects can increase. The proportion of female participation in Non-technical subjects with a result higher or equal to 3.50(table 16 ; Group 2 (x2)), 0.27 is the average and the standard deviation is 0.052. Group 2 (x2) is extensively spread out with respect to its average. Hence, female participation in Non-technical subjects with a result less than 3.00 and greater than or equal to 2.50 have a small-scale increasing rate. Thereby, Group 1(x1) has the possibility of increasing the rate of participation on technical subjects more than the participation in non-technical subjects with a result less than 3.00 and greater than or equal to 2.50.

Result Female in Technical Vs Non Technical fields Category 3

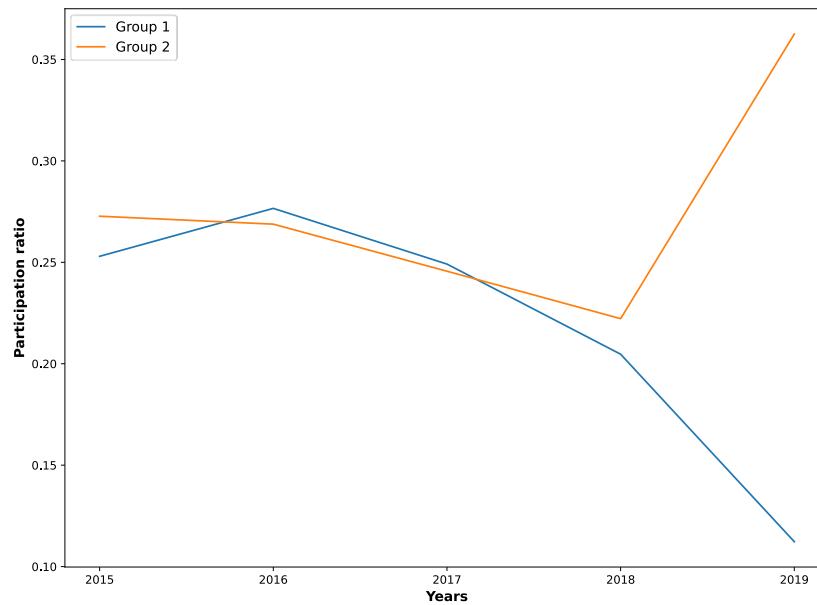


Figure 18 : Female participation in Technical and Non-technical Subjects Category 3

Table 17 and Figure 18 deal with category 3 which represents the below-average results. So, the lower is better in this case. Here Table 17 and Figure 18 show that the participation of group 2 is more than group 1. In 2016 and 2017 the result is different. The average difference is 0.28 which means that on average group 2 participated more than group 1 which is a direct indicator that group 2 did worse than group 1. And in 2019 the difference is noticeable.

So, it is clear that in all three categories on average group 1 did better than group 2 although the total enrolment and yearly enrolment ratio are higher for group 2.

Experiment 3

Here, this study focuses on Male vs Female results In Technical Subjects for 5 years.

Category 1: It focuses on the participation ratio of male and female students with a result higher or equal to 3.50.

Table 18: Male vs Female participation ratio for Category 1

Year	Group 1 (x1)	Group 3(x2)	Difference (x2-x1)
2015	0.23	0.19	-0.04
2016	0.23	0.2	-0.03
2017	0.29	0.25	-0.04
2018	0.33	0.27	-0.06
2019	0.43	0.33	-0.1
Average	0.3	0.25	-0.05
S.D	0.083	0.056	0.028

In proportion to male students with a result higher or equal to 3.50 (Table 18 ; Group 1 (x1)), the standard deviation is 0.083 and the average is 0.3. The yearly ratio of data is marginally spread out with respect to the average as the standard deviation is 0.083. The yearly ratio data are close to each data [100]according to the average. Nevertheless, Group 1 (x1) data consistency is less than Group 2 (x2).

Regarding the participation of females with a result higher or equal to 3.50 (Table 18 ; Group 2 (x2)), 0.25 is the average and the standard deviation is 0.056. Group 2 (x2) is extensively spread out with respect to its average. Hence, female participation with a result higher or equal to 3.50 has a small-scale increasing rate. However, female participation increased gradually. The consistency of Group 2 (x2) data is more than Group 1 (x1). Accordingly, Female participation is far better than male participants in technical subjects.

Male vs Female Result in Technical fields Category 1

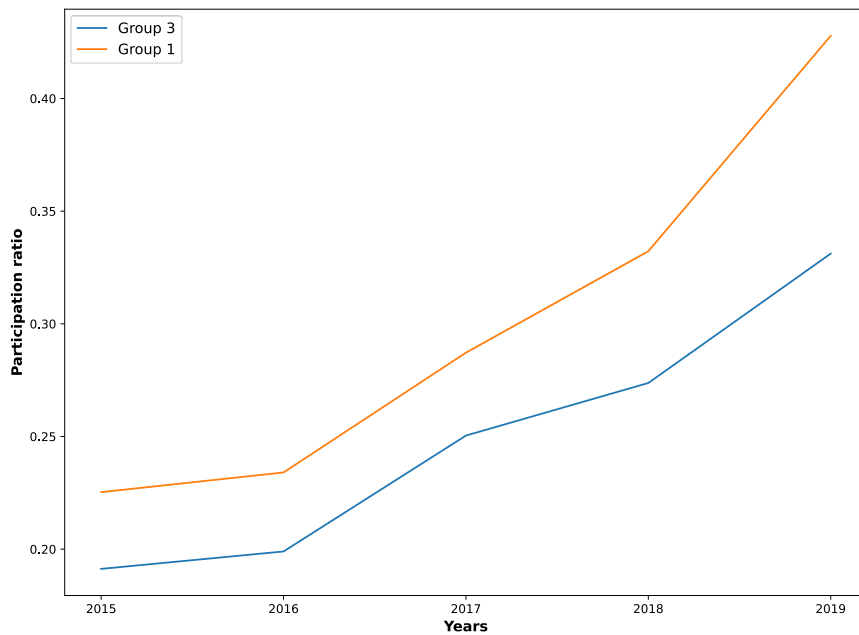


Figure 19 : Male and Female participation ratio in Technical Subjects for Category 1

Table 17 and Figure 19 show that group 1 is doing better than group 3. The difference column shows that the result difference is increasing the difference was lowest in 2016-2017 but in 2019 it increased to 0.1 which is the highest and the figure shows the same each year the difference is negative which means group 1 is doing better than group 3. Here, the average difference is -0.126 which shows on average that group 1 is doing better. The result of five years shows that the result participation in both groups is increasing every year. This means both males and females are doing better in technical subjects. But the female ratio is higher so Females are doing better than their male counterparts.

Category 2: It focuses on the participation ratio of male and female students with a result of less than 3.50 and greater than or equal to 3.00.

Table 19: Male vs Female participation ratio in Technical Subjects for Category 2

Year	Group 1 (x1)	Group 2(x2)	Difference (x2-x1)
2015	0.34	0.35	0.01
2016	0.38	0.35	-0.03
2017	0.36	0.34	-0.02
2018	0.37	0.36	-0.01
2019	0.44	0.44	0.00
Average	0.38	0.37	-0.01
S.D	0.037	0.040	0.016

With respect to male students with a result less than 3.50 and greater than or equal to 3.00 (Table 18; Group 1 (x1)), the standard deviation is 0.037 and the average is 0.38. The yearly ratio of data is widely spread out with respect to the average. The yearly ratio data are not so close to each data [100] according to the average.

Concerning the participation of females with a result less than 3.50 and greater than or equal to 3.00 (Table 19; Group 2 (x2)), 0.37 is the average and the standard deviation is 0.040. Group 2 (x2) is not more demonstrated with respect to its average. Thereby, female participation with a result less than 3.50 and greater than or equal to 3.00 has a growing rate. The female participation ratio in technical subjects is more than the male participation ratio in technical subjects.

Male vs Female Result in Technical fields Category 2

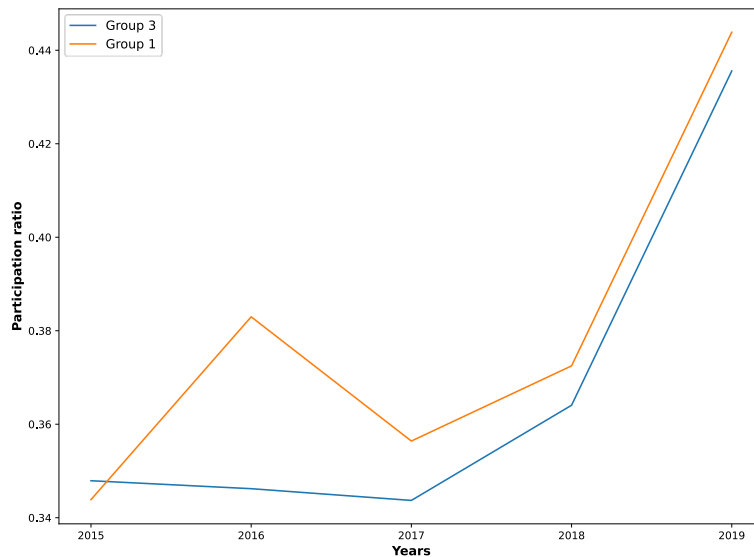


Figure 20 : Male and Female participation ratio in Technical Subject for Category 2

Table 19 and Figure 20 for Male vs female performance in category 2 shows some interesting findings. In 2015 female students' performance was less than male students but in 2016 it changed and showed a huge jump from 0.34 to 0.38, on the other hand, the male performance remain nearly constant till 2017 with values between 0.35-0.34. Although there is a slight decline in the female result in 2016-2017 still it is higher than the male students. In 2018-2019 the male and female results showed the same level of incline. But the increase in the male result is better than the increase in the female result. Here the average difference is -0.01 which means the average difference is very small but still female students outperformed the male students.

Category 3: It focuses on the participation ratio of male and female students with a result less than 3.00 and greater than or equal to 2.50.

Table 20 : Male and Female participation ratio in Technical Subject for Category 3

Year	Group 1 (x1)	Group 3 (x2)	Difference(x2-x1)
2015	0.25	0.28	-0.03
2016	0.28	0.29	0.01
2017	0.25	0.26	0.01
2018	0.2	0.25	0.05
2019	0.11	0.21	0.10
Average	0.22	0.26	0.03
S.D	0.066	0.031	0.049

With regard to male students with a result less than 3.00 and greater than or equal to 2.50 (Table 20 ; Group 1 (x1)), the standard deviation is 0.066 and the average is 0.22. The yearly ratio of data is narrowly spread out in respect of average. The yearly ratio data are near to each data in respect to average.

As for the participation of females with a result less than 3.00 and greater than or equal to 2.50 (Table 19; Group 2 (x2)), 0.26 is the average and the standard deviation is 0.031. Group 2 (x2) is more demonstrated with respect to its average. Therefore, the Male participation ratio in technical subjects is more than the female participation ratio in technical subjects with a result less than 3.00 and greater than or equal to 2.50.

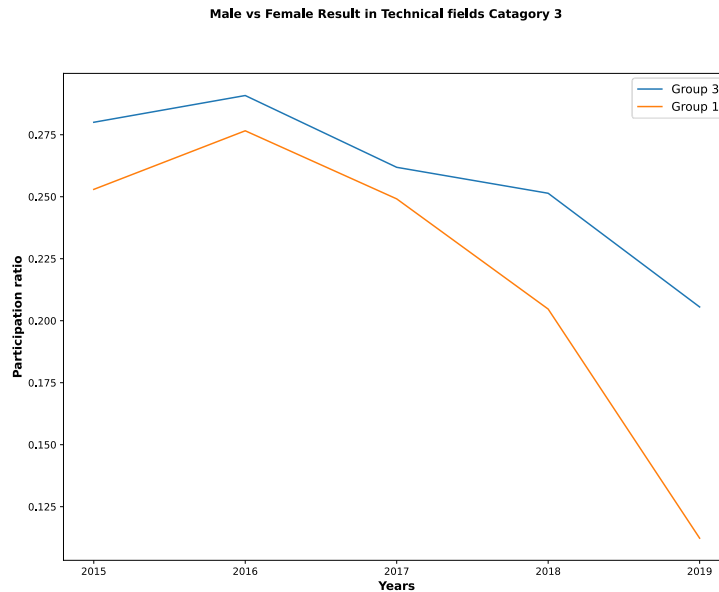


Figure 21: Male and Female participation ratio in Technical Subject for Category 2

In category 3 Table 20 and Figure 21 show that the contribution of both groups is decreasing. But the rate of change of group 1 is higher than group 3. The from the figure is clear that the decline of group 1 is higher as the ‘Group 1’ line is downward and stiffer than the ‘Group 3’ line.

Experiment 4

In this experiment, the aim is to find out the contribution of female students in good results (category 1) with respect to total students in technical fields.

Table 21 : Female contribution in the good result

Year	Total	Female	Contribution in %
2015	0.42	0.23	54.76%
2016	0.43	0.23	53.48%
2017	0.54	0.29	53.70%
2018	0.61	0.33	54.09%
2019	0.76	0.43	56.57%
Average	0.55	0.3	54.52
S.D	0.140	0.083	1.244

To sum up, table 21 shows the total students' contribution in technical fields to get good results and female students' contribution in technical fields to get a good result. According to the concept of standard deviation and average females are the good result holder. For total standard deviation is 0.140 and the average is 0.55. On the other hand, for female participants, the standard deviation is 0.083 and the average is 0.3. The total student's deviation has a 0.41 difference from the average. Female students' standard deviation has 0.217 difference from average. This expresses that the total student's data strew more than female student's data ($0.41 > 0.217$) [100]. So female students contributed more than total students.

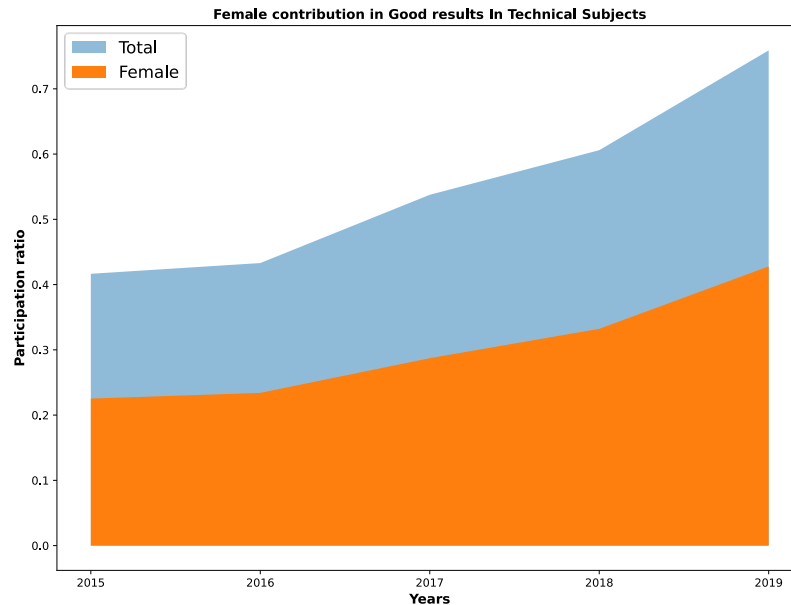


Figure 22: Female Contribution in the good result

Table 21 and Figure 22 show more than 50% of group 1 are contributing to good results here performance remains mostly constant and group 1 outperformed group 3. In 2016 and 2017 the contribution gets lower but it increases in 2018 and 2019 and reached its peak with an approximate contribution of 57%. And the contribution of group 3 is the opposite here.

So, it is clear that female students are contributing more than male students in good results.

4.2 Discussion

Some studies were conducted on text classification and female studies. Nevertheless, no study was conducted for tagging for sorting paper to find out the actual domain and also female contribution in STEM. The study of tagging was organized using information from over 9 million data. we have used 5 different methods and classifiers in this study. According to the result of experimental studies, SVM gives better results on content base tagging comparing journal base tagging. This study of the “Classification and Recommendation system” uses NB, LR, SVM, Conv Net, and ANN were selected. The

experiments are conducted on Article Title, Author Keyword, and Abstract. 10 different experiments were done on the datasets to prepare the base models and 3 more experiments to measure the aggregated ensemble model. For better interpretation, the experiments will be clustered/packed in sets.

Female contribution in STEM was conducted using 8999 students. 4 experiment was conducted on a result dataset of five years (2015-2019) of female students in female education. According to the result of experimental studies, female students are doing better in both technical and non-technical fields rather than females.

Chapter 5

IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY

5.1 Impact on Society

In our experiment, it is clear that content-based tagging methods are better than journal-based tagging and provide a better correlation between tags and contents. So, content-based tagging methods will allow authors or readers to obtain desired and related information faster. This can contribute to society by maintaining time and resources.

The tool presented in this paper can be used to check research paper domains. It can help authors check and understand the domains their paper is targeting and modify the papers as they desire to make them more correlated to the desired domain. If the paper is written in the intended field, the reader will have no trouble finding the appropriate paper, which will be to everyone's advantage.

This study also shows the bias against female students in the STEM and technical fields is nothing but ill-founded rumors. On the other hand, it shows that in the technical subjects in academia, females are performing as well as their counterparts. Females are basic to capture the potential that technology and sustainability offer our future and society. Consequently, teaching female students will advance culture.

5.2 Ethical Aspects

We attempted to demonstrate two forms of tagging in our first section and which is preferable to employ. If the author used machine learning to write their paper, then content-based tagging should be used to publish it. Publishing that material in a journal for the sole purpose of the publication is unethical. The author should publish in accordance with his paper type to give the readers what they deserve.

In our second part, we focused on ethics by making sure the reader wasn't defrauded in any way. Consider a person who reads a publication and decides they want to study computer science but finds that it primarily focuses on topics related to medical computer science.

As a result, the researcher wasted their time, and the author received a negative review. The author made a mistake by not properly identifying and abstracting their work. This is detrimental to our culture. Because there are occasions when reading a paper costs money. The reader will not find this morally acceptable if it keeps happening. As a result, the writer will comprehend their paper and create appropriate tags for it using our machine learning technique.

Male students have always gotten greater attention in their field of study. People believe teaching women is a waste of time and resources. They thus started to ignore the female student. Human rights are not served by this. Equal opportunities for men and women are essential. According to our research, female students do better than male students when given enough resources and equal opportunities. Society will become more conscious as a result of this research.

5.3 Sustainability Plan

A project sustainability plan must be created before the project starts. Sustainability ensure the outputs, outcomes, and benefits. A sustainability plan mainly gives us a realistic picture of how a project will run and what the project's future plans are. The goal of our tool is to assist authors to find out the domain of the paper correctly. Authors published papers in different domains and they sometimes fail to achieve the main domain. This model must be simple for the author to adjust.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE

6.1 Summary of the Study

In order to determine which had a higher association between tags and contents, we employed a tool in our study report. To determine whether content-based tagging or journal-based tagging is more effective, we employed machine learning algorithms like Naive Bayes, Logistic regression, Support vector machines, CNN, and ANN. According to the findings, content-based tagging techniques offer a greater association between tags and contents than journal-based tagging. For our second section, we developed a tool to assist authors in determining the domains that their papers are targeting and in making any necessary modifications to their papers so that they are better aligned with the targeted domain. In our final part, we have demonstrated that preconceptions about female students in STEM and technical professions are unfounded rumors. It demonstrates that women are performing as well to their male counterparts in academic technical topics.

6.2 Conclusions

This study is divided into three distinct areas. The process of sorting the papers is known as tagging. Different methods are used by publishers to sort the papers. A "Descriptive Research" of two different data tagging techniques, namely content-based tagging, and journal-based tagging, makes up the first section of the paper. This experiment was conducted in 2 phases, each with 6 experiments on 3 types of data, namely the "Article Title, Abstract, and Author Keyword". So, in each phase, 18 experiments are done on different input types with an 80-20 split where 80% of data is used for training and 20% for testing. The first phase deals with WOS categories and the second phase deals with Dimension Categories. In each phase, there were 2 types of algorithms applied. Traditional machine Learning and Deep learning. For the first part Naïve Bayes, Logistic Regression,

Random Forest, Support Vector Machine, And for the second part Naïve Bayes, Logistic Regression, and Support Vector Machine from the traditional machine learning algorithms has been used. The second section, titled "Applied research," is where we put the information from the first portion to use and develop a mechanism for categorizing research publications. It can be seen that while experimenting with abstract and author keywords the ensemble algorithm achieves more than 6% and 4% accuracy in both Dimension and WOS datasets respectively. The increase in accuracy on article titles is not as significant as the previous 2 but still, the improvement is more than 4% and 3% in both datasets. From the result it is clear that content-based tagging methods are better than journal-based tagging and provide a better correlation between tags and contents. The third and last section is called "Exploratory Research," where we looked for gender bias in STEM and technical fields regarding academic males and females. Here we have found that total female participation in technical subjects is 1262 which is only 17% of total participation. And in Non-technical subjects' participation is 459 which is 33.30%. But each year the ratio increased considerably. On average, only 15% more females enrolled in the non-technical subjects and in 2017 it was increased by 21.05%. The female contribution to a successful outcome was lower in 2016 and 2017 but increased in 2018 and 2019, reaching its high with an about 57% contribution. It is clear that the number of females in technical subjects is doing better and is increasing each year from 2015 to 2019 as the performance curve is upward sloping. The study on female participation in STEM involved 8999 students. On a result dataset of five years (2015-2019) of female students in female education, four experiments were carried out. According to the findings of experimental investigations, female students do better than male students in both technical and non-technical fields.

6.3 Limitations

In text classification tasks we only targeted a few classes. An array of new classes can be studied. Newer transformer-based methods such as BERT and others can be implemented. Different numbers of features can be selected and different methods of vectorizers can be

used. In ensemble methods, the base algorithms can be improved to have an overall improved result. Different Feature engineering methods can be used.

In the Gender study, this paper focuses on only the descriptive analysis but a regression analysis is also possible. On the other hand, It only focuses on tertiary-level education.

6.4 Implication for Further Study

In the future, we will try to expand our work on these projects of text classification and female contribution in STEM. We will find a way to get a better result in our research. In the future, we will try to improve the performance of the model that we have already used in our research. We will apply extra methods for expanding our study.

APPENDIX

Abbreviation:

NB = Naïve Bayes

LR = Logistic Regression

RF = Random Forest

SVM = Support Vector Machine

CNN = Convolutional Neural Networks

ANN = Artificial Neural Networks

STEM = Science, Technology, Engineering, and Math

Appendices: Research Reflection:

We had very limited knowledge about ensemble models and STEM when we started this research project. Our supervisor was a kind person. He always helped us when we were willing to help. He provided us with inestimable advice that was quite helpful for us. He guided us as an advisor. His advice assisted us to complete this project. He is very calm towards us. We learned how to manage and many more new things from our supervisor. There were some problems in the first step as we worked with over 9 million instances in text classification and also worked on female contributions in education. But we recovered all the problems now. Finally, Conducting the research gave us the courage and motivated us to do more research in the future.

Reference

- [1] S. Klasen, "The Impact of Gender Inequality on Economic Performance in Developing Countries," <https://doi.org/10.1146/annurev-resource-100517-023429>, vol. 10, pp. 279–298, Oct. 2018, doi: 10.1146/ANNUREV-RESOURCE-100517-023429.
- [2] M. A. Andersson and C. E. Harnois, "Higher exposure, lower vulnerability? The curious case of education, gender discrimination, and Women's health," *Soc Sci Med*, vol. 246, p. 112780, Feb. 2020, doi: 10.1016/J.SOCSCIMED.2019.112780.
- [3] A. Gaye, J. Klugman, W. Bank, and E. Zambrano, "Measuring Key Disparities in Human Development: The Gender Inequality Index The Social Institutions and Gender Index: A Reformulation View project Multidimensional Deprivation-Measurement View project," 2010, Accessed: Jan. 02, 2022. [Online]. Available: <https://www.researchgate.net/publication/254419513>
- [4] J. Erikson and C. Josefsson, "Does Higher Education Matter for MPs in their Parliamentary Work? Evidence from the Swedish Parliament," <https://doi.org/10.1080/00344893.2019.1581077>, vol. 55, no. 1, pp. 65–80, Jan. 2019, doi: 10.1080/00344893.2019.1581077.
- [5] R. Ahmed and N. Hyndman-Rizk, "The higher education paradox: towards improving women's empowerment, agency development and labour force participation in Bangladesh," <https://doi.org/10.1080/09540253.2018.1471452>, vol. 32, no. 4, pp. 447–465, May 2018, doi: 10.1080/09540253.2018.1471452.
- [6] A. Almiaçık, F. Gökşen, and D. Yüksek, "School to work or school to home? An analysis of women's vocational education in Turkey as a path to employment," <https://doi.org/10.1080/09540253.2018.1465897>, vol. 31, no. 8, pp. 1040–1056, Nov. 2018, doi: 10.1080/09540253.2018.1465897.
- [7] S. Sinha Mukherjee, "More educated and more equal? A comparative analysis of female education and employment in Japan, China and India," <https://doi.org/10.1080/09540253.2015.1103367>, vol. 27, no. 7, pp. 846–870, Nov. 2015, doi: 10.1080/09540253.2015.1103367.
- [8] P. Shah and A. Khurshid, "Muslim womanhood, education, and empowerment: ethnographic reflections from Pakistan and India," <https://doi.org/10.1080/09540253.2018.1543859>, vol. 31, no. 4, pp. 458–474, May 2019, doi: 10.1080/09540253.2018.1543859.
- [9] L. Smylie, E. Maticka-Tyndale, and D. Boyd, "Adolescent Sexual Health Planning Committee. Evaluation of a school-based sex education programme delivered to Grade Nine students in Canada," *Sex Educ*, vol. 8, no. 1, pp. 25–46, Feb. 2008, doi: 10.1080/14681810701811795.
- [10] D. Kirby, "Understanding what works and what doesn't in reducing adolescent sexual risk taking," *Fam Plan Perspect*, vol. 33, no. 6, pp. 276–81, 2001, doi: 10.2307/3030195.
- [11] K. P. Phillips and A. Martinez, "Sexual and Reproductive Health Education: Contrasting Teachers', Health Partners' and Former Students' Perspectives," *Canadian Journal of Public Health* 2010 101:5, vol. 101, no. 5, pp. 374–379, Sep. 2010, doi: 10.1007/BF03404856.
- [12] H. P. Schaalma, C. Abraham, M. R. Gillmore, and G. Kok, "Sex education as health promotion: What does it take?," *Arch Sex Behavior*, vol. 33, no. 3, pp. 259–69, Jun. 2004, doi: 10.1023/b:aseb.0000026625.65171.1d.
- [13] C. Goldin, "The Quiet Revolution That Transformed Women's Employment, Education, and Family".
- [14] J. Ehrlinger and D. Dunning, "How Chronic Self-views Influence (and Potentially Mislead) Estimates of Performance," *J Pers Soc Psychol*, vol. 84, no. 1, pp. 5–17, 2003, doi: 10.1037/0022-3514.84.1.5.
- [15] S. J. Correll, "Gender and the Career Choice Process: The Role of Biased Self-Assessments1," <https://doi.org/10.1086/321299>, vol. 10, no. 6, pp. 1691–1730, Jul. 2015, doi: 10.1086/321299.
- [16] C. Schuster and S. E. Martiny, "Not Feeling Good in STEM: Effects of Stereotype Activation and Anticipated Affect on Women's Career Aspirations," *Sex Roles*, vol. 76, no. 1–2, pp. 40–55, Jan. 2017, doi: 10.1007/S11199-016-0665-3.
- [17] "The Global Gender Gap Report 2017 | World Economic Forum." <https://www.weforum.org/reports/the-global-gender-gap-report-2017> (accessed Dec. 01, 2021).

- [18] A. Hussénius, “Trouble the gap: gendered inequities in STEM education,” <https://doi.org/10.1080/09540253.2020.1775168>, vol. 32, no. 5, pp. 573–576, Jul. 2020, doi: 10.1080/09540253.2020.1775168.
- [19] Catherine. Hill, Christianne. Corbett, Andresse. st. Rose, and American Association of University Women., “Why So Few? Women in Science, Technology, Engineering, and Mathematics.,” American Association of University Women, p. 109, 2010.
- [20] D. N. Beede, T. A. Julian, D. Langdon, G. McKittrick, B. Khan, and M. E. Doms, “Women in STEM: A Gender Gap to Innovation,” STEM (Science, Technology, Engineering, and Mathematics) Workforce Trends and Policy Considerations, pp. 51–61, Aug. 2011, doi: 10.2139/SSRN.1964782.
- [21] A. M. C. L. C. S. I. Munoz Boudet, “Women and STEM in Europe and Central Asia. Report No: AUS0002179.,” World Bank, 2021, Accessed: Dec. 29, 2021. [Online]. Available: www.worldbank.org
- [22] S. Cheryan, J. O. Siy, M. Vichayapai, B. J. Drury, and S. Kim, “Do Female and Male Role Models Who Embody STEM Stereotypes Hinder Women’s Anticipated Success in STEM?;,” <http://dx.doi.org/10.1177/1948550611405218>, vol. 2, no. 6, pp. 656–664, Apr. 2011, doi: 10.1177/1948550611405218.
- [23] S. H. Keng, “Gender bias and statistical discrimination against female instructors in student evaluations of teaching,” *Labour Econ*, vol. 66, p. 101889, Oct. 2020, doi: 10.1016/J.LABECO.2020.101889.
- [24] “Girls’ Education Overview.” <https://www.worldbank.org/en/topic/girlseducation> (accessed Dec. 01, 2021).
- [25] “Primary education, pupils (% female) - Pakistan, Bangladesh, India, Indonesia, Nepal, China | Data.” <https://data.worldbank.org/indicator/SE.PRM.ENRL.FE.ZS?end=2019&locations=PK-BD-IN-ID-NP-CN&start=1970&view=chart> (accessed Jan. 03, 2022).
- [26] “Secondary education, pupils (% female) - Pakistan, Bangladesh, India, Indonesia, Nepal, China | Data.” <https://data.worldbank.org/indicator/SE.SEC.ENRL.FE.ZS?end=2019&locations=PK-BD-IN-ID-NP-CN&start=2010> (accessed Jan. 03, 2022).
- [27] “Girls’ Education Overview.” <https://www.worldbank.org/en/topic/girlseducation#1> (accessed Jan. 03, 2022).
- [28] A. Salatino, F. Osborne, and E. Motta, “CSO Classifier 3.0: a scalable unsupervised method for classifying documents in terms of research topics,” *International Journal on Digital Libraries*, vol. 23, no. 1, pp. 91–110, Mar. 2022, doi: 10.1007/S00799-021-00305-Y/FIGURES/5.
- [29] M. Daradkeh, L. Abualigah, S. Atalla, and W. Mansoor, “Scientometric Analysis and Classification of Research Using Convolutional Neural Networks: A Case Study in Data Science and Analytics,” *Electronics* 2022, Vol. 11, Page 2066, vol. 11, no. 13, p. 2066, Jun. 2022, doi: 10.3390/ELECTRONICS11132066.
- [30] C. Caragea, J. Wu, S. das Gollapalli, and C. L. Giles, “Document Type Classification in Online Digital Libraries”, Accessed: Dec. 21, 2022. [Online]. Available: <http://pdfbox.apache.org/>
- [31] B. Kandimalla, S. Rohatgi, J. Wu, and C. L. Giles, “Large Scale Subject Category Classification of Scholarly Papers With Deep Attentive Neural Networks,” *Front Res Metr Anal*, vol. 5, p. 31, Feb. 2021, doi: 10.3389/FRMA.2020.600382/BIBTEX.
- [32] F. Hoppe, D. Dessi, and H. Sack, “Deep Learning meets Knowledge Graphs for Scholarly Data Classification,” *The Web Conference 2021 - Companion of the World Wide Web Conference, WWW 2021*, pp. 417–421, Apr. 2021, doi: 10.1145/3442442.3451361.
- [33] G. Piao, “Scholarly Text Classification with Sentence BERT and Entity Embeddings,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12705 LNAI, pp. 79–87, 2021, doi: 10.1007/978-3-030-75015-2_8/COVER.
- [34] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, “Deep Learning--based Text Classification,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, Apr. 2021, doi: 10.1145/3439726.
- [35] B. Kandimalla, S. Rohatgi, J. Wu, and C. L. Giles, “Large Scale Subject Category Classification of Scholarly Papers With Deep Attentive Neural Networks,” *Front Res Metr Anal*, vol. 5, p. 31, Feb. 2021, doi: 10.3389/FRMA.2020.600382/BIBTEX.

- [36] G. Maillette De Buy Wenniger, T. van Dongen, E. Aedmaa, H. T. Kruitbosch, E. A. Valentijn, and L. Schomaker, "Structure-Tags Improve Text Classification for Scholarly Document Quality Prediction", doi: 10.18653/v1/P17.
- [37] A. Mccallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification".
- [38] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu, "Transferring Naive Bayes Classifiers for Text Classification", Accessed: Nov. 06, 2022. [Online]. Available: www.aaai.org
- [39] E. Frank and R. R. Bouckaert, "Naive bayes for text classification with unbalanced classes," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4213 LNAI, pp. 503–510, 2006, doi: 10.1007/11871637_49/COVER.
- [40] W. Zhang and F. Gao, "An Improvement to Naive Bayes for Text Classification," *Procedia Eng.*, vol. 15, pp. 2160–2164, Jan. 2011, doi: 10.1016/J.PROENG.2011.08.404.
- [41] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," <https://doi.org/10.1177/0165551516677946>, vol. 44, no. 1, pp. 48–59, Nov. 2016, doi: 10.1177/0165551516677946.
- [42] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," *2019 International Conference on Automation, Computational and Technology Management, ICACTM 2019*, pp. 593–596, Apr. 2019, doi: 10.1109/ICACTM.2019.8776800.
- [43] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augmented Human Research 2020 5:1*, vol. 5, no. 1, pp. 1–16, Mar. 2020, doi: 10.1007/S41133-020-00032-0.
- [44] C. N. Kamath, S. S. Bukhari, and A. Dengel, "Comparative study between traditional machine learning and deep learning approaches for text classification," *Proceedings of the ACM Symposium on Document Engineering 2018, DocEng 2018*, vol. 18, Aug. 2018, doi: 10.1145/3209280.3209526.
- [45] O. M. Aborisade and M. Anwar, "Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers," *Proceedings - 2018 IEEE 19th International Conference on Information Reuse and Integration for Data Science, IRI 2018*, pp. 269–276, Aug. 2018, doi: 10.1109/IRI.2018.00049.
- [46] Y. Chen and Z. Zhang, "Research on text sentiment analysis based on CNNs and SVM," *Proceedings of the 13th IEEE Conference on Industrial Electronics and Applications, ICIEA 2018*, pp. 2731–2734, Jun. 2018, doi: 10.1109/ICIEA.2018.8398173.
- [47] X. Luo, "Efficient English text classification using selected Machine Learning Techniques," *Alexandria Engineering Journal*, vol. 60, no. 3, pp. 3401–3409, Jun. 2021, doi: 10.1016/J.AEJ.2021.02.009.
- [48] K. Zeng, Z. Pan, Y. Xu, Y. Q.-J. M. Informatics, and undefined 2020, "An ensemble learning strategy for eligibility criteria text classification for clinical trial recruitment: algorithm development and validation," medinform.jmir.org, Accessed: Nov. 10, 2022. [Online]. Available: <https://medinform.jmir.org/2020/7/e17832>
- [49] G. Wang, J. Sun, J. Ma, K. Xu, J. G.-D. support systems, and undefined 2014, "Sentiment classification: The contribution of ensemble learning," Elsevier, Accessed: Nov. 10, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167923613001978>
- [50] D. Liang, B. Y.-I. Sciences, and undefined 2021, "Two-stage three-way enhanced technique for ensemble learning in inclusive policy text classification," Elsevier, Accessed: Nov. 10, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0020025520308112>
- [51] D. K.-C. B. U. J. of Science and undefined 2016, "The effect of ensemble learning models on Turkish text classification," dergipark.org.tr, vol. 12, pp. 215–220, doi: 10.18466/cbujos.04526.
- [52] X. Dong, Z. Yu, W. Cao, Y. Shi, Q. M.-F. of C. Science, and undefined 2020, "A survey on ensemble learning," Springer, vol. 2020, no. 2, pp. 241–258, Apr. 2020, doi: 10.1007/s11704-019-8208-z.
- [53] L. Shi, X. Ma, L. Xi, Q. Duan, J. Z.-E. S. with Applications, and undefined 2011, "Rough set and ensemble learning based semi-supervised algorithm for text classification," Elsevier, Accessed: Nov. 10, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417410013072>

- [54] A. Onan, "Hybrid supervised clustering based ensemble scheme for text classification," *Kybernetes*, vol. 46, no. 2, pp. 330–348, 2017, doi: 10.1108/K-10-2016-0300/FULL/HTML.
- [55] A. R.-P. of the 18th C. on Scientometrics and undefined 2021, "Protégé-advisor gender-pairings in academic survival and productivity of German PhD graduates," *researchgate.net*, Accessed: Nov. 09, 2022. [Online]. Available: https://www.researchgate.net/profile/Andreas-Rehs/publication/354131016_Protege-advisor_gender_pairings_in_academic_survival_and_productivity_of_German_PhD_graduates/links/61268fab9659a41297d9cb70/Protege-advisor-gender-pairings-in-academic-survival-and-productivity-of-German-PhD-graduates.pdf
- [56] M. Nikunen, K. L.-G. and Education, and undefined 2020, "Gendered strategies of mobility and academic career," *Taylor & Francis*, vol. 32, no. 4, pp. 554–571, May 2018, doi: 10.1080/09540253.2018.1533917.
- [57] S. Cohen, P. Hanna, J. Higham, D. Hopkins, and C. Orchiston, "Gender discourses in academic mobility," *Gend Work Organ*, vol. 27, no. 2, pp. 149–165, Mar. 2020, doi: 10.1111/GWAO.12413.
- [58] H. Jöns, "Transnational academic mobility and gender," *Globalisation, Societies and Education*, vol. 9, no. 2, pp. 183–209, Jun. 2011, doi: 10.1080/14767724.2011.577199.
- [59] C. Thun, "Excellent and gender equal? Academic motherhood and 'gender blindness' in Norwegian academia," *Gend Work Organ*, vol. 27, no. 2, pp. 166–180, Mar. 2020, doi: 10.1111/GWAO.12368.
- [60] H. Boekhout, I. van der Weijden, L. W. preprint arXiv, and undefined 2021, "Gender differences in scientific careers: A large-scale bibliometric analysis," *arxiv.org*, Accessed: Nov. 09, 2022. [Online]. Available: <https://arxiv.org/abs/2106.12624>
- [61] H. Boekhout, I. van der Weijden, L. W. preprint arXiv, and undefined 2021, "Gender differences in scientific careers: A large-scale bibliometric analysis," *arxiv.org*, Accessed: Nov. 09, 2022. [Online]. Available: <https://arxiv.org/abs/2106.12624>
- [62] K. Z.- Sociologica and undefined 2011, "How gender neutral are state policies on science and international mobility of academics?," *rivisteweb.it*, Accessed: Nov. 09, 2022. [Online]. Available: <https://www.rivisteweb.it/doi/10.2383/34631>
- [63] G. Nández and Á. Borrego, "Use of social networks for academic purposes: A case study," *Electronic Library*, vol. 31, no. 6, pp. 781–791, 2013, doi: 10.1108/EL-03-2012-0031/FULL/HTML.
- [64] S. Haustein, R. Costas, and V. Larivière, "Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns," *PLoS One*, vol. 10, no. 3, Mar. 2015, doi: 10.1371/JOURNAL.PONE.0120495.
- [65] S. Klar, Y. Krupnikov, J. B. Ryan, K. Searles, and Y. Shmargad, "Using social media to promote academic research: Identifying the benefits of twitter for sharing academic work," *PLoS One*, vol. 15, no. 4, 2020, doi: 10.1371/JOURNAL.PONE.0229446.
- [66] X. Zhang, D. Wen, J. Liang, and J. Lei, "How the public uses social media wechat to obtain health information in China: A survey study," *BMC Med Inform Decis Mak*, vol. 17, Jul. 2017, doi: 10.1186/S12911-017-0470-0.
- [67] S. Banshal, V. Singh, P. Muhuri, P. M. preprint arXiv, and undefined 2019, "How much research output from India gets social media attention?," *arxiv.org*, Accessed: Nov. 09, 2022. [Online]. Available: <https://arxiv.org/abs/1909.03506>
- [68] S. Banshal, V. Singh, P. M.-O. I. Review, and undefined 2021, "Can altmetric mentions predict later citations? A test of validity on data from ResearchGate and three social media platforms," *emerald.com*, Accessed: Nov. 09, 2022. [Online]. Available: <https://www.emerald.com/insight/content/doi/10.1108/OIR-11-2019-0364/full/Altmetric.com>
- [69] S. K. Banshal, V. K. Singh, and P. K. Muhuri, "Can altmetric mentions predict later citations? A test of validity on data from ResearchGate and three social media platforms," *Online Information Review*, vol. 45, no. 3, pp. 517–536, 2020, doi: 10.1108/OIR-11-2019-0364/FULL/ALTMETRIC.COM.
- [70] S. Banshal, A. Basu, V. Singh, ... S. G. preprint arXiv, and undefined 2020, "Do 'altmetric mentions' follow Power Laws? Evidence from social media mention data in Altmetric. com," *arxiv.org*, Accessed: Nov. 09, 2022. [Online]. Available: <https://arxiv.org/abs/2011.09079>

- [71] A. Nath and S. Jana, "An Altmetric Analysis of Scholarly Publications from Earth and Planetary Science Discipline: An Exploratory Study of Indian Publications," researchgate.net, Accessed: Nov. 09, 2022. [Online]. Available: https://www.researchgate.net/profile/Amit-Nath-3/publication/360862883_An_Altmetric_Analysis_of_Scholarly_Publications_from_Earth_and_Planetary_Science_Discipline_An_Exploratory_Study_of_Indian_Publications/links/62922e8bc660ab61f84cddc7/An-Altmetric-Analysis-of-Scholarly-Publications-from-Earth-and-Planetary-Science-Discipline-An-Exploratory-Study-of-Indian-Publications.pdf
- [72] S. Banshal, V. Singh, P. Muhuri, P. M. preprint arXiv, and undefined 2019, "Disciplinary variations in altmetric coverage of scholarly articles," arxiv.org, Accessed: Nov. 09, 2022. [Online]. Available: <https://arxiv.org/abs/1910.04205>
- [73] W. Huang, P. Wang, and Q. Wu, "A correlation comparison between Altmetric Attention Scores and citations for six PLOS journals," PLoS One, vol. 13, no. 4, Apr. 2018, doi: 10.1371/JOURNAL.PONE.0194962.
- [74] S. Banshal, V. Singh, ... G. K.-J. of I., and undefined 2018, "An altmetric analysis of scholarly articles from India," content.iospress.com, Accessed: Nov. 09, 2022. [Online]. Available: <https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs169495>
- [75] C. Meschede, T. S.- Scientometrics, and undefined 2018, "Cross-metric compatability and inconsistencies of altmetrics," Springer, Accessed: Nov. 09, 2022. [Online]. Available: <https://link.springer.com/article/10.1007/s11192-018-2674-1>
- [76] S. Banshal, S. Gupta, H. Lathabai, V. S.-J. of Informetrics, and undefined 2022, "Power Laws in altmetrics: An empirical analysis," Elsevier, Accessed: Nov. 09, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S175115772200061X>
- [77] M. Anderson et al., "Russell Sage Foundation)," 2013.
- [78] F. P. Byrne et al., "Tools and techniques for solvent selection: green solvent selection guides," Sustainable Chemical Processes 2016 4:1, vol. 4, no. 1, pp. 1–24, May 2016, doi: 10.1186/S40508-016-0051-Z.
- [79] B. Ertl, S. Luttenberger, and M. Paechter, "The impact of gender stereotypes on the self-concept of female students in STEM subjects with an under-representation of females," Front Psychol, vol. 8, no. MAY, p. 703, May 2017, doi: 10.3389/FPSYG.2017.00703/BIBTEX.
- [80] K. R. Thorson, C. E. Forbes, A. B. Magerman, and T. v. West, "Under threat but engaged: Stereotype threat leads women to engage with female but not male partners in math," Contemp Educ Psychol, vol. 58, pp. 243–259, Jul. 2019, doi: 10.1016/J.CEDPSYCH.2019.03.012.
- [81] W. M. Williams, "Editorial: Underrepresentation of women in science: International and cross-disciplinary evidence and debate," Front Psychol, vol. 8, no. JAN, p. 2352, Jan. 2018, doi: 10.3389/FPSYG.2017.02352/BIBTEX.
- [82] R. Singh, Y. Zhang, M. (Maggie) Wan, and N. A. Fouad, "Why do women engineers leave the engineering profession? The roles of work–family conflict, occupational commitment, and perceived organizational support," Hum Resour Manage, vol. 57, no. 4, pp. 901–914, Jul. 2018, doi: 10.1002/HRM.21900.
- [83] L. Perez-Felkner, S. Nix, and K. Thomas, "Gendered pathways: How mathematics ability beliefs shape secondary and postsecondary course and degree field choices," Front Psychol, vol. 8, no. MAR, pp. 386–386, Apr. 2017, doi: 10.3389/FPSYG.2017.00386/FULL.
- [84] D. I. Miller, K. M. Nolla, A. H. Eagly, and D. H. Uttal, "The Development of Children's Gender-Science Stereotypes: A Meta-analysis of 5 Decades of U.S. Draw-A-Scientist Studies," Child Dev, vol. 89, no. 6, pp. 1943–1955, Nov. 2018, doi: 10.1111/CDEV.13039.
- [85] M. Frank Fox, D. G. Johnson, and S. v Rosser, "WCIMI N, (d NO 1~, A N D TEC HNOLOGY."
- [86] S. L. Hacker, "The culture of engineering: Woman, workplace and machine," Womens Stud Int Q, vol. 4, no. 3, pp. 341–353, Jan. 1981, doi: 10.1016/S0148-0685(81)96559-3.
- [87] J. S. McIlwee and J. G. Robinson, Women in engineering: Gender, power, and workplace culture. SUNY Press, 1992.
- [88] E. Makarova, B. Aeschlimann, and W. Herzog, "The Gender Gap in STEM Fields: The Impact of the Gender Stereotype of Math and Science on Secondary Students' Career Aspirations," Front Educ (Lausanne), vol. 4, p. 60, Jul. 2019, doi: 10.3389/FEDUC.2019.00060/BIBTEX.

- [89] B. A. Nosek et al., “National differences in gender–science stereotypes predict national sex differences in science and math achievement,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10593–10597, Jun. 2009, doi: 10.1073/PNAS.0809921106.
- [90] T. Petersen and L. A. Morgan, “Separate and Unequal: Occupation-Establishment Sex Segregation and the Gender Wage Gap,” <https://doi.org/10.1086/230727>, vol. 101, no. 2, pp. 329–365, Oct. 2015, doi: 10.1086/230727.
- [91] M. Broer, Y. Bai, and F. Fonseca, “A Review of the Literature on Socioeconomic Status and Educational Achievement,” *IEA Research for Education*, vol. 5, pp. 7–17, 2019, doi: 10.1007/978-3-030-11991-1_2.
- [92] W. H. Kim and J. Lee, “The Effect of Accommodation on Academic Performance of College Students With Disabilities,” *Rehabil Couns Bull*, vol. 60, no. 1, pp. 40–50, Oct. 2016, doi: 10.1177/0034355215605259.
- [93] H. Naher, T. Tanim, and N. Sultana, “Women in Science and Technology: A study in Bangladesh,” *Sociology and Anthropology*, vol. 7, no. 7, pp. 306–312, Aug. 2019, doi: 10.13189/SA.2019.070702.
- [94] J. Paswan and V. K. Singh, “Gender and research publishing analyzed through the lenses of discipline, institution types, impact and international collaboration: a case study from India,” *Scientometrics* 2020 123:1, vol. 123, no. 1, pp. 497–515, Feb. 2020, doi: 10.1007/S11192-020-03398-5.
- [95] M. Shoaib and H. Ullah, “Female and Male Students’ Educational Performance in Tertiary Education in the Punjab Pakistan,” 2019, Accessed: Dec. 11, 2021. [Online]. Available: <https://www.researchgate.net/publication/338621831>
- [96] D. Gabriel, “Race, ethnicity and gendered educational intersections,” <https://doi.org/10.1080/09540253.2021.1967667>, vol. 33, no. 7, pp. 791–797, 2021, doi: 10.1080/09540253.2021.1967667.
- [97] D. Fokum, D. Coore, Y. L.-F.-P. of the 47th ACM, and undefined 2016, “The performance of female computer science students across three Caribbean Islands,” *dl.acm.org*, pp. 419–424, Feb. 2016, doi: 10.1145/2839509.2844566.
- [98] “pickle5 · PyPI.” <https://pypi.org/project/pickle5/> (accessed Nov. 10, 2022).
- [99] “sklearn.svm.LinearSVC — scikit-learn 1.1.3 documentation.” <https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html?highlight=svm> (accessed Oct. 31, 2022).
- [100] R. Delmas and Y. Liu, “EXPLORING STUDENTS’ CONCEPTIONS OF THE STANDARD DEVIATION 8”, Accessed: Feb. 27, 2022. [Online]. Available: <http://www.stat.auckland.ac.nz/serj>

Invalid Index

ORIGINALITY REPORT

10%

SIMILARITY INDEX

7%

INTERNET SOURCES

6%

PUBLICATIONS

4%

STUDENT PAPERS

PRIMARY SOURCES

1

Md. Abdullah-Al-Kafi, Israt Jahan Tasnova, Md. Wadud Islam, Sumit Kumar Banshal. "Chapter 53 Performances of Different Approaches for Fake News Classification: An Analytical Study", Springer Science and Business Media LLC, 2022

Publication

3%

2

dspace.daffodilvarsity.edu.bd:8080

Internet Source

1%

3

www.textanalytics.in

Internet Source

1%

4

journal.acs.org.au

Internet Source

1%