BANGLA NEWS ARTICLE CATEGORIZATION USING MACHINE LEARNING

BY

MD AL SHAHRIAR HAQUE
ID: 191-15-2646
AND

UMME SHAWDA
ID: 191-15-2457

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

Supervised By

Mohammad Jahangir Alam
Sr. Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Amit Chakraborty Chhoton
Sr. Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project/internship titled **“Bangla news article categorization using Machine learning”**, submitted by Md Al Shahriar Haque and Umme Shawda , ID No: 191-15-2646, 191-15-2457 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment 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 *30/01/2023*.

<u>BOARD OF EXAMINERS</u>

Chairman

____

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

SArefin
____

Dr. Mohammad Shamsul Arefin
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

____

Ms. Sharmin Akter
Lecturer (Senior Scale)
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

____

Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Mohammad Jahangir Alam, Sr. Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Jahangir 30.01.23

Mohammad Jahangir Alam
Sr. Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Amit Chakraborty Chhoton
Sr. Lecturer
Department of CSE
Daffodil International University

Submitted by:

(Md Al shahriar Haque)
ID: -191-15-2646
Department of CSE
Daffodil International University

Umme shawda

(Umme shawda)
ID: -191-15-2457
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

We really grateful and wish our profound our indebtedness to **Mohammad Jahangir Alam**,**Amit Chakraborty Chhoton** Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "**Bangla news article categorization using 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 draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Professor Dr. Touhid Bhuiyan**, **Professor & Head,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

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

ABSTRACT

Bangla language got familiar many years ago in the world and Many online Bangla news portals are growing day by day. We can get news within a few seconds with their help of them. Some media are telecasting news by live stream and some are publishing news through online news portals. With their help of them, much news is being published day by day. We are familiar with many new things by seeing/reading the news. This news is not separated by its specific categories the problem arrives because every people don't like every category. For this reason, they feel disturbed to read the news but very few researchers are working in Bangla news and at this time data gap is increasing very rapidly. In this paper, we try to solve this problem by Machine learning. we collect data by the web crawler. Our dataset has 408470 rows and collects data 120 thousand. We use label mapping for category labeling and to get sequence we use a tokenizer, for data preprocessing we use a slicer to get the same sample in every category. We use flatten, embedding, and dense, and we use 'adam' optimizer, for loss function 'sparse categorical cross entropy', for visualizations we use a heatmap, and confusion matrix, for classification we use some classifiers like SVM, KNN, decision tree, random forest, naive bayes, GradientBoostingClassifier. After using the decision tree, and random forest we get a training accuracy is 98.08%.

TABLE OF CONTENTS

CONTENTS **PAGE**

CHAPTER

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction

The Internet makes our life more comfortable. We can get news from one country to another country within a few moments. At this time Bangla language is rapidly spreading in many countries. Many news channels are working to publish the Bangla language like BBC, Prothom Alo, Jugantor, Ittefaq, etc. Much data are gathered in our Bangla treasury but very few researchers work in the Bangla language[7]. In our research project, our main goal is to collect many Bangla news data and implement it for those who are feeling bored reading newspapers Because in our newspaper many categories of news are found on one page. Every man doesn't like every category of news, so All news are being categorized by the given headline[28] in this time if a man wants to read a specific category of news then it is possible and many people are interested to read Bangla newspapers. Python has a library that is Scrapy, by using Scrapy we select the Prothom Alo newspaper URL and our requested column data is (Author, published data, title, category, and content). To start crawling we select a date that is 2013 to 2021 in JSON format.After stopping crawling we converted it JSON to CSV by python library that is JSON, CSV. Because our dataset is very big that is not possible to handle our computer. Our dataset has 10,14000 rows. After that, we find the unique category, in our dataset has 9 categories and label mapping by 0,1,2,3. . . that is specific for all categories. we use a tokenizer to get the sequence in the text. In preprocessing we use Slicer in a category and every category data sample is 5000 and reduces dimension from many to single dimension. For modeling, we use Embedding, Flatten, Dense. For training, we use the Adam optimizer and the loss function is calculated by sparse categorical cross-entropy. In visualization, we use a Confusion matrix and heat map. For classification, we use some classifiers like(Random forest, decision tree, SVM, KNN, and Gradient Boosting). At this time Random forest and decision tree give the best accuracy which is 98.08. Our system can accurately predict the news headline.

Table 1.1.1 : Our sample Dataset .

author	category	published_date	tag	title	content
গাজীপুর প্রতিনিধি	bangladesh	০৪ জুলাই ২০১৩, ২৩:২৬	গাজীপুর	কালিয়াকৈরে টিফিন খেয়ে ৫০০ শ্রমিক অসুস্থ, বিক্ষোভ	গাজীপুরের কালিয়াকৈর উপজেলার তেলিরচালা এলাকায় আজ বৃহস্পতিবার রাতের টিফিন খেয়ে একটি পোশাক কারখানার ৫০০ শ্রমিক অসুস্থ হয়ে পড়েছেন। এ ঘটনায় বিক্ষোভ করেছেন ওই কারখানার
অনলাইন ডেস্ক	sports	০৪ জুলাই ২০১৩, ২৩:০৯	টেনিস	সেমিফাইনাল বাধাও পেরিয়ে গেলেন লিসিকি	এবারের উইম্বলডনটা স্মরণীয় করে রাখার মিশনেই যেন নেমেছেন সাবিনা লিসিকি। চতুর্থ রাউন্ডের লড়াইয়ে সেরেনা উইলিয়ামসকে হারিয়ে শুরু করেছিলেন স্বপ্নযাত্রা। কোয়ার্টার ফাইনালে কাইয়া কানেপিকে হারাতে খুব একটা বেগ পেতে হয়নি। তবে সেমিফাইনালে কঠিন প্রতিপক্ষের মুখেই পড়তে হয়েছিল লিসিকিকে।
অনলাইন ডেস্ক	technology	০৪ জুলাই ২০১৩, ২১:৩৭	গবেষণা	পাসওয়ার্ড ভুলে যান!	সহজ পাসওয়ার্ডের কারণে অনলাইন অ্যাকাউন্ট সহজেই হ্যাক হয়ে যেতে পারে। জটিল পাসওয়ার্ড মনে রাখা কষ্টকর। তবে জটিল পাসওয়ার্ড মনে রাখার কষ্টের দিন হয়তো ফুরাল। পাসওয়ার্ড সহজে মনে রাখার ঝামেলা থেকে মুক্তি দিতে বিকল্প অনেক উপায় নিয়ে কাজ করছেন গবেষকেরা।
অনলাইন ডেস্ক	entertainment	০৪ জুলাই ২০১৩, ২১:০৫	হলিউড	পুনর্মিলনের আশায় কেটিকে ক্রুজের চিঠি	বিচ্ছেদের এক বছর পরও কেটি হোমসকে ভুলতে পারছেন না ‘মিশন ইমপসিবল’ তারকা টম ক্রুজ। আর তাই তো সম্প্রতি তিনি সাবেক স্ত্রীকে আবেগঘন এক চিঠি লিখেছেন পুনর্মিলনের আশাবাদ ব্যক্ত করে

1.2 Motivation

Social media news is not separated by its specific categories as a result problem has arrived because people don’t like every category. For this reason, they are feeling disturbed to read news. We will try accurate preprocessing, try to improve accuracy, and add more categories. Many works have already been done on this related topic but they only work on Machine learning and their data amount is very low. In our research, we have collected a huge dataset that is of large volume. Our collected data amount is around 4 lakhs, on the other hand, ours have many categories like(author, content, details, tag, title, news time, etc). Maximum researchers work only on English news article categorization but we are trying to do Bangla news article categorization. I hope from this work of ours there will be a great amount of work for Bangla news. Besides, I hope we will get a good outcome from our research.

1.3 Rationale of the Study

In our research paper, we will use some basic Python libraries for data collection and data processing. As all our data will be collected from the ‘Prothom Alo’ Bangla newspaper. That would be very difficult for us to collect manually, so we will use Python’s Scrapy library. To use this library a researcher needs to learn the full process of the web crawler. They also need to learn Deep learning's tokenizer. For data handling, we will need to use JSON format data and as that amount is very high, it is impossible to handle in any other way. So a researcher must know how to handle huge data. A researcher must also learn some Deep learning features like as dense, flatten, embedding, etc.

1.4 Research Questions

- Which are the challenges faced by Bangla newspaper article categorization applications?
- How does Bangla newspaper article categorization Impacts on our everyday life?
- Which approach is better for Bangla newspaper article categorization?
- How Can we apply machine learning and deep learning models?

1.5 Expected Output

In this paper, we will try to work on Bangla and social media news categorization using Deep learning. We will categorize all news by its headline and get which category of news it will be. Our target will be to collect more data and process them and get accurate results by applying specific models.

1.6 Project Management and Finance

In our paper, we have two authors and our supervisor and co-supervisor to help us. In this paper, the first author has contributed to the coding implementation and paper writing and the second author's contribution is to writing the paper and formatting. During the whole time, our supervisor helped us and always gave us direction. We didn't get any financial help from any organization.

1.7 Report Layout

In our report, there are a total of six chapters where we have discussed and explained all our work in as organized a way as possible. Below, we are giving a very short overview for understanding it better.

Chapter 1

We have covered all our introduction, motivation of work, our objectives, and research related questions in this chapter. We have discussed the theory and work-related information before and after our research.

Chapter 2

In this section, we have discussed related work for Bangla and English QA systems. Also, we covered the research summary and scope of our research problem. At the end of this chapter, we covered challenges we have faced during this research.

Chapter 3

In this chapter, we explained the methodology of our work. Also covered the required technologies and equipment we have used. We have shown some sample data and the source for English and Bangla QA both. Lastly, we covered data pre-processing, statistical analysis, model description, and representation of the Taxonomy of our model.

Chapter 4

We have discussed results and shown the comparison between Bangla and English data graphs. We have shown real-life prediction accuracy using tables for both of the languages we have worked on.

Chapter 5

We covered chapter five by the impact on society, ethical aspects, and sustainability plan to understand the importance of our research work.

Chapter 6

Finally, we covered and discussed future work. We have explained the summary of the study, recommendation, conclusion, and implication for further study.

CHAPTER 2

Background

2.1 Terminologies

Research background can have some terminology such as Abstract, Data, Variable, Concept, Sample, Assumption, Population, Hypothesis, Construct, etc. All of the topics must be covered in successful research.

2.2 Related Work

Our research topic is "News categorization by Natural language processing". We have read approximately 34 research papers from google scholar and Researchgate. Maximum research papers are published in IEEE conferences. We want to make a system that will filter or categorize from any context or headline[10]. Because most people don't need to read all types of categories, as every person has their interests. There are many document categorization systems developed for English language processing but there is no usable system developed for Bangla texts[2]. They have used spec2 vector, word embedding, for classification word embedding[6], SGD, and multiclass SVM and they have collected data from the web, blogs newspapers, and online books. Their research document has around 14k+ data. And their testing accuracy is 93%[3]. Their future work is to gather more classes and add more documents for the best accuracy[4]. University of Dhaka CSE department published an article where they had collected a dataset of around 3,76,226 data. They have used tokenizer and word 2 vec. And their accuracy is around 95%. In the future, they are trying to work on NLP-related problems and want to use CNN and LSTM[1]. Besides, they are also considering improving the prediction model. Another team also wants to reduce the Bangla data gap but in that research, they will use CNN[16] available in Sci_kit learn, count vectorization and at the same time they will use N_gram for text slicing. Some similar articles have been published. Their main goal is that since most people dislike seeing all types of news from the newspaper or social media[26] so they want to make a system that can easily identify the user preference. A team from Daffodil International university has already done on Bangla newspaper. They have used N-gram for text categorization. Maximum researchers use naive Bayes, SVM, Decision tree, and Random Forest[8]. In the future, they will collect more data to get better accuracy. We have collected another language’s research paper which is Urdu news classification[20] using machine learning but we don't have any wish to do another research on other languages like Chinese or Japanese[9]. A research team at BUET university wants to make a tool where everybody can give context or headline and it will also predict a class. A research team of Stony Brock university USA has worked with word embedding[24]. They want to make a cluster that also uses another platform like document classification[18], sentiment analysis, parts-of-speech tagging, named entity recognition and machine translation, etc. They have also worked on a huge dataset of around 5,19,20,010 data. Their predicted feature is 6, for clustering they used a K-means cluster and its accuracy is 94%. In the future, they want to collect more data because if the data volume is huge, then the testing accuracy will increase. A research team of BRAC university they have worked on N-gram news classification only[13]. They have used Zipf’s Law[33]. They recommend that from their experiment they have seen that character-level trigram perform better than any other N-grams[32]. At Chitkara University, Himachal Pradesh, India there also have been working on Neural networks[14]. Where they used stop word removal feature Selection(Boolean weighting, Class Frequency Thresholding, Term Frequency Inverse Class Frequency, Information Gain.)[34].

2.3 Comparative Analysis and Summary

A team from Daffodil International university has already done on Bangla newspaper. They have used N-gram for text categorization[29]. Maximum researchers use naive Bayes[30], SVM, Decision tree, and Random Forest[5]. In the future, they will collect more data to get better accuracy. We have collected another language’s research paper which is Urdu news classification using machine learning but we don't have any wish to do another research on other languages like Chinese or Japanese. A research team at BUET university wants to make a tool where everybody can give context or headline and it will also predict a class. A research team of Stony Brock university USA has worked with word embedding. They want to make a cluster that also uses another platform like document classification, sentiment analysis, parts-of-speech tagging, named entity recognition and machine translation, etc. They have also worked on a huge dataset of around 5,19,20,010 data. Their predicted feature is 6, for clustering, they used a K-means cluster and its accuracy is 94%. In the future, they want to collect more data because if the data volume is huge, then the testing accuracy will increase. A research team of BRAC university they have worked on N-gram news classification only[11]. They have used Zipf's Law. They recommend that from their experiment they have seen that character-level trigram perform better than any other N-grams. At Chitkara University, Himachal Pradesh, India there also have been working on Neural networks[18]. Where they used stop word removal feature Selection(Boolean weighting, Class Frequency Thresholding, Term Frequency Inverse Class Frequency, Information Gain.).

2.4 Scope of the Problem

There are many document categorization system developed for English language processing but there is no usable system developed for Bangla texts. The most common problem is to categorize the textual documents. Document classification has paramount importance on several applications like searching, filtering, and organizing the textual documents. That's why we work with Bangla newspaper article categorization.

2.5 Challenges

In our research paper, our main challenge will be to collect data. As our data is Bangla News related we can't do any kind of request to another organization or anywhere. We can collect data in two ways manually and we can also collect data by python library but not everyone can easily implement it. Implementing a web crawler is more challenging. Our data type is textual and textual data type preprocessing is very challenging. Data pre-processing is more difficult because as the dataset gets huge we can not use it in any local software like Jupyter notebook or Spyder or any platform. After all, it will show runtime errors or memory errors. As a result, the data reading part will be more challenging. As such applying a huge data model will be more challenging in our research and also produce good output.

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

Our research topic is Bangla News Categorization. These news are not separated by their specific categories as a result problem arrives because each people doesn’t like every category. For this reason, they are feeling disturbed to read the news but very few researchers are working in Bangla news and at this time data gap is increasing very rapidly. In this paper, we will try to solve this problem by using Neural network[21]. For Data collection we need web crawler. For crawling data from Bengali online newspapers, we need the python “Scrapy” library. For data preprocessing we will use slicer. That’s why we have to use natural language processing. And we made a model by embedding, flatten, Dense. We will use 75 percent of the data for training and 25 percent for testing. To predict Bangla News Categorization in our research we use some Fig. 2. model classifiers like Support Vector Machine (SVM), Decision Tree, Random Forest, KNN and Gradient boosting[15].

3.2 Data Collection Procedure/Dataset Utilized

We have collected data by web crawler using Python’s “Scrapy” library. At this time we have collected minimum 12 lakhs data but we are using only 4 lakhs. In the future we will use all our collected data. Our dataset has 10 columns and in the category column we have 9 unique categories (bangladesh, economy, education, entertainment, international, life style, opinion, sports , technology).

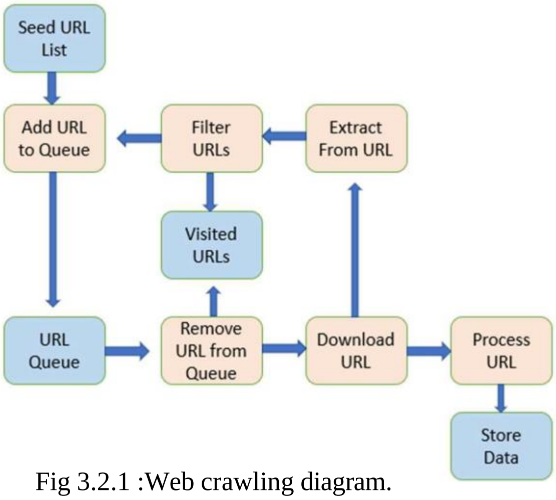
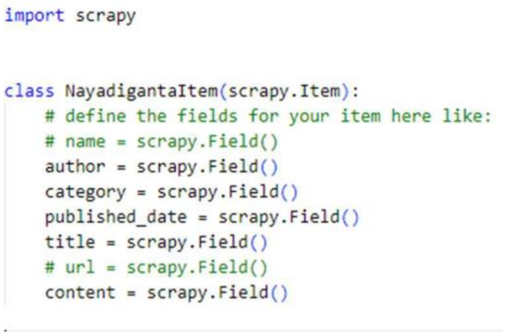


Fig 3.2.1 :Web crawling diagram.

At first we need to seed URL, after that wee need to add URL inside the Queue. We can use Scrapy here for picking some item that we want to add inside the Queue . After that we put those items that we have cracked from URL’s inside the fields. For defining the fields for our item an example code is :

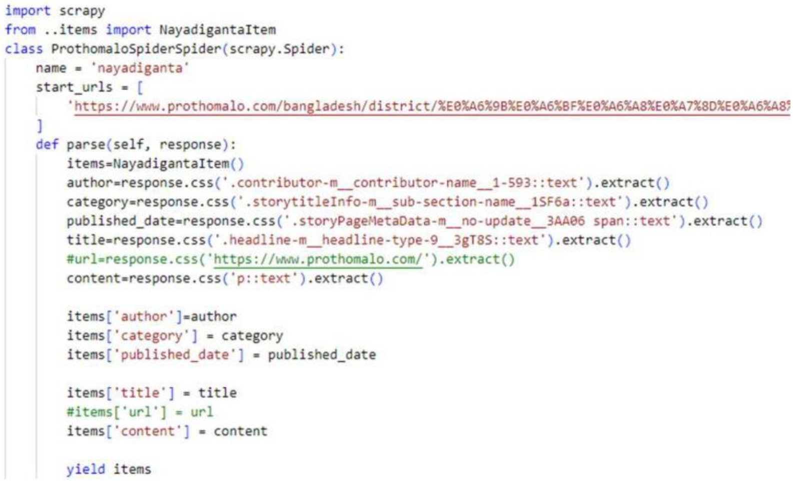


```
import scrapy


class NayadigantaItem(scrapy.Item):
    # define the fields for your item here like:
    # name = scrapy.Field()
    author = scrapy.Field()
    category = scrapy.Field()
    published_date = scrapy.Field()
    title = scrapy.Field()
    # url = scrapy.Field()
    content = scrapy.Field()
```

Fig 3.2.2 :scrapy those items that we need.

For getting data we need to specify those attributes that we want to put it in those items in scrapy. We can take help from a Spider ,a spider process is given in Fig-3.2.2



```
import scrapy
from ..items import NayadigantaItem
class ProthomaloSpiderSpider(scrapy.Spider):
    name = 'nayadiganta'
    start_urls = [
        'https://www.prothomalo.com/bangladesh/district/%E0%A6%9B%E0%A6%BF%E0%A6%A8%E0%A7%8D%E0%A6%A8
    ]
    def parse(self, response):
        items=NayadigantaItem()
        author=response.css('.contributor-m__contributor-name__1-593::text').extract()
        category=response.css('.storytitleInfo-m__sub-section-name__1SF6a::text').extract()
        published_date=response.css('.storyPageMetaData-m__no-update__3AA06 span::text').extract()
        title=response.css('.headline-m__headline-type-9__3gT8S::text').extract()
        #url=response.css('https://www.prothomalo.com/').extract()
        content=response.css('p::text').extract()

        items['author']=author
        items['category'] = category
        items['published_date'] = published_date

        items['title'] = title
        #items['url'] = url
        items['content'] = content

        yield items
```

Fig 3.2.2 :spider data crawling way and put it in scrapy items.

At first we need start URL then define the function that take items from spider that is shown in Fig-3.2.2. Secondly we need extract response CSS that we can find by inspecting the site. We get it by inspect individual elements. Lastly, we put the CSS in the Scrapy items .

3.3 Scope of the Problem

Data collection is the major part of research but data collection and noise cancelation is very difficult . When we take tag data from news portal then we face problem that is many unnecessary tag columns like: tag0,tag1,tag2 etc but all contains null value. It is very difficult to detect where the actual value is.

3.4 Statistical Analysis

In our research we have collected huge Data, its amount around 12 lakhs for our ram limitation we have used 4 lakhs only. We have unique 9 categories of data like as bangladesh, economy, education, entertainment, international, life style , opinion , sports , technology all data details are given below.

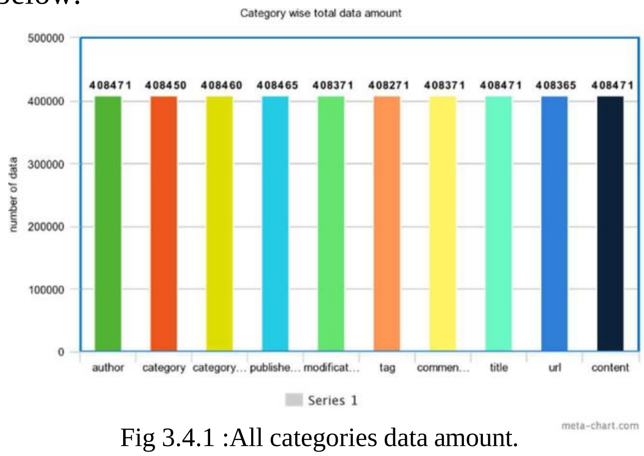


Fig 3.4.1 :All categories data amount.

3.5 Proposed Methodology

Data collection is an essential issue to research and nothing is exceptional in our case. We need to collect data. After collect data we will preprocess our dataset.

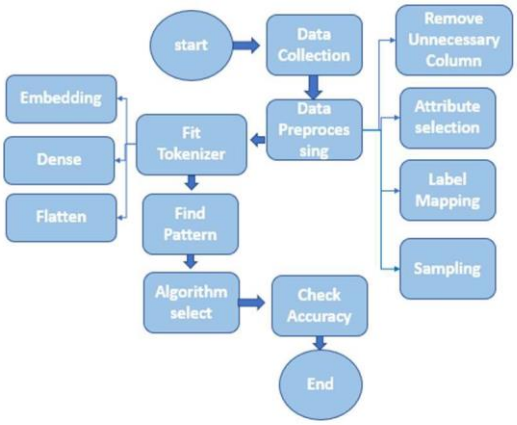


Fig 3.5.1 :Proposed methodology.

We have to remove html tag, remove some unnecessary columns for all the news crawled from Bengali online newspapers. For data preprocess we use slicer to get the same sample in every category. That’s why we have to use natural language processing. And we made a model by embedding flatten . In this fig:01 firstly we have to input data in an input layer. Suppose that our data set have a title (“ ”) after inputting this title and passing input layer this title have to face the embedding system [14]. Embedding system is natural language processing where word embedding is a term used for the representation of words for text analysis and stores these words so that it can reuse these words for use as needed later [15]. After finishing the title embedding this data will go to the process Flatten. The concept of flatten data is data where data from related database tables or flat file records are gathered into a single or reduced number of tables [16]. By destroying a big data table it will create many small tables according to its characteristics. This reverses the process of normalization where data is organized so that each fact is stored once and avoids the duplication of data. The next step is dense data. Dense data can be described as many different pieces of the required information on a specific kind of a subject, no matter whatever the subject happens to be. Data can be dense data many times for its needs. We use 75 percents data for training and 25 percents data for testing . To predict Bangla News Categorization in our research we use some Fig. 2. modeling classifiers like Support Vector Machine (SVM) , Decision Tree,Random Forest, KNN Gradient boosting. [17]

3.6 Implementation Requirements

IN this paper we are try to solve this problem by neural network. For Data collection we need web crawler. For crawling data from Bengali online newspapers we need python “Scrapy” library. For data preprocess we use slicer. That’s why we have to use natural language processing. And we made a model by embedding , flatten , Dense . In this fig:01 firstly we have to input data in an input layer. Suppose that our data set have a title (“ ”) after inputting this title and passing input layer this title have to face the embedding system [14]. Embedding system is natural language processing where word embedding is a term used for the representation of words for text analysis and stores these words so that it can reuse these words for use as needed later [15]. After finishing the title embedding this data will go to the process Flatten. The concept of flatten data is data where data from related database tables or flat file records are gathered into a single or reduced number of tables [16]. By destroying a big data table it will create many small tables according to its characteristics. This reverses the process of normalization where data is organized so that each fact is stored once and avoids the duplication of data. The next step is dense data. Dense data can be described as many different pieces of the required information on a specific kind of a subject, no matter whatever the subject happens to be. Data can be dense data many times for its needs. Finally getting performance we need some model like as Vector Machine (SVM) , Decision Tree,Random Forest, KNN Gradient boosting[17].

CHAPTER 4

Machine LearningAlgorithm

4.1 Support Vector Machine (SVM)

For the Bangla News Categorization task, we use a Support Vector Machine (SVM) classification algorithm. [5] SVM is a popular supervised learning algorithm for classification tasks and many researchers attempted to perform document classification tasks using it [18]. Given a set of training documents, each document is marked with a particular category. The proposed methodology advocates analytic parameter selection directly from the training data, rather than re-sampling approaches commonly used in SVM applications[12]. SVM is Effective in high-dimensional spaces. In cases where the the number of dimensions is greater than the number of samples, there also we can use SVM [19]. But in our case, we cannot get good accuracy. SVM has some common parts such as support vector, Hyperplane, Marginal distance, Linear separable, Non-linear separable. All of these are given below.

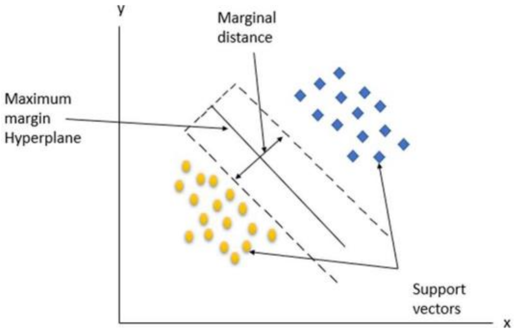


Fig 4.1.1 : SVM Hyperplane.

Here SVM detect that, which data point is closest the marginal line.

4.2 Decision Tree

In Machine Learning where the data is continuously split according to a certain parameter. And this can explain what the input is and what the corresponding output is in the training data set. Higher predictive precision can usually be acquired by generating multiple trees from the data, all of which are used in categorizing a new model. More than one test can be used to partition the models at each stage, giving families of superimposed trees, or multiple training sets can be samples from the data. The predictions from several trees can be connected by simple voting or by more cultivated techniques such as piling. Specifically, in decision analysis operations research we use Decision Tree to help identify a strategy most likely to reach the goal. Decision Tree gives the best Training Accuracy.

$$Entropy(S) = 1 - \sum_{i=1}^{n} -p_i \, log_2 \, p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{vE \text{ values}} \frac{|s_v|}{|s|} \times Entropy(s_v)$$

In Decision tree we need to calculate the Gain and try to improve gain score. Before calculate it firstly we calculate each Entropy.

4.3 Random Forest

If we want to solve a problem that combines many classifiers into complex problems then, we have to use a Random Forest [22]. Because utilizing ensemble learning it can provide a solution [23]. A random forest is a Machine learning technique that's used to solve regression and classification problems. and other tasks that operate by constructing a multitude of decision trees at training time [24]. For classification tasks, the output of the random forest is the class selected by most trees. Random decision forests correct for decision trees' habit of overfitting to their training set. By using these classifiers we got good training accuracy [25]

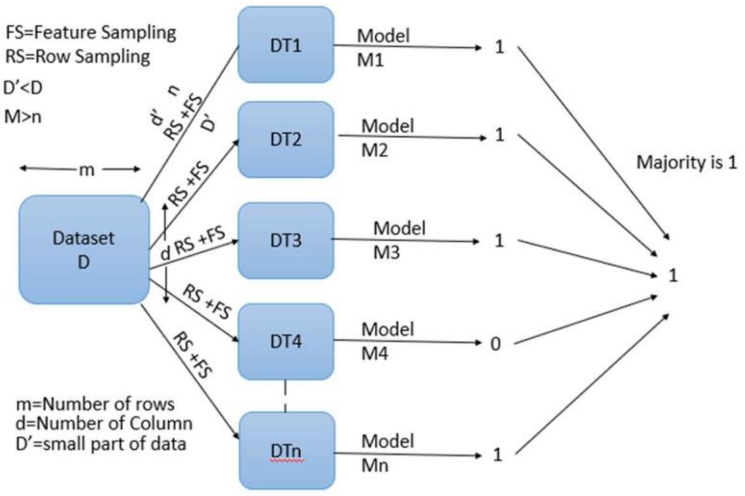


Fig 4.3.1 : Random forest working procedure.

In Fig 4.3.2, we try to describe that we have a big dataset D. It has m number of rows and d numbers of the column. A small number of data perform in a single Decision Tree algorithm with row sampling and feature sampling. After that Decision tree makes a model for binary classification means 0 and 1. In the last stage full system chooses the majority.

4.4 k-nearest neighbors (KNN)

KNN means The k-nearest neighbors (KNN) algorithm. This is a simple, supervised Machine learning algorithm [2]. This is a non-parametric classification method. It can be used to solve both classification and regression problems. KNN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation.

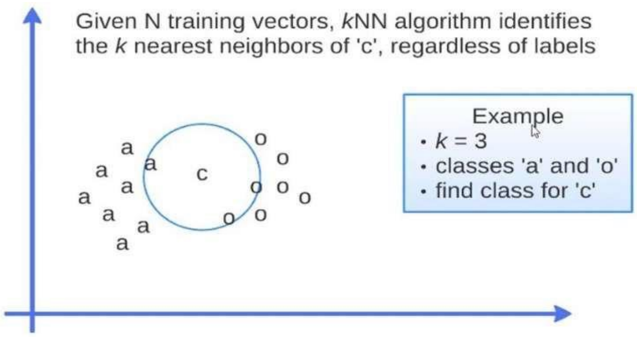

Given N training vectors, kNN algorithm identifies the k nearest neighbors of 'c', regardless of labels
a
c
o
Example
• k = 3
• classes 'a' and 'o'
• find class for 'c'

Fig 4.4.1 : KNN working procedure.

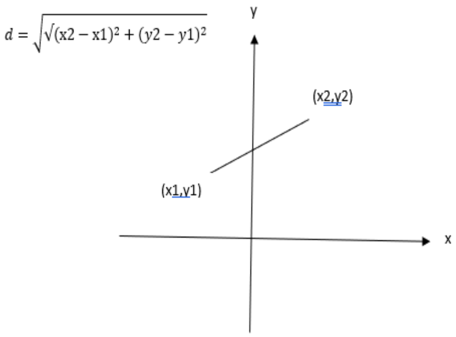


Fig 4.4.2 : Euclidean Distance .

In Fig 4.4.1 we see the KNN working procedure. KNN follows some steps. The first step is select the value of k which means the nearest value. The second step is to calculate the Euclidean Distance by the law that is shown in Fig 4.4.2, it will find each data point distance. Thirdly, take the K nearest neighbors as per the calculated Euclidean Distance. Fourthly, among these K neighbors, count the number of the data points in each category. Finally, assigns the new data points to that category for which the number of neighbor is maximum. Our model will be ready.

CHAPTER 5

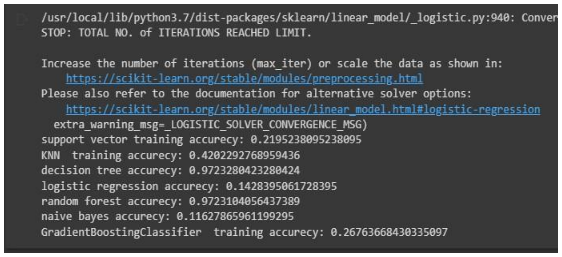
Experimental Result and Discussion

5.1 Experimental Setup

In our research, we have used some models to get the result of which algorithm gives us the better result we can use those algorithms for categorization. In our experiment Decision tree, and Random forest gives us better output on the other hand SVM and KNN are not performing better than those algorithms we will see all compare in the Discussion section.

5.2 Experimental Results & Analysis

In our research, we use some classifiers like SVM, decision tree, random forest, KNN, and Gradient boosting, at this time random forest, and decision tree give the best accuracy of all classifiers. Random forest gives 97.23 percent and decision tree gives 97.23 percent. But at this time another classifier is not predicting very good performance. All results are shown in fig-5. In the confusion matrix test accuracy is 66.27 percent. Show in Table 1. Our All confusion matrices test result in fig-5.2.2.

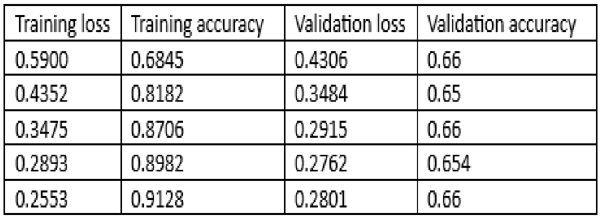


```
/usr/local/lib/python3.7/dist-packages/sklearn/linear_model/_logistic.py:940: Conver
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG)
support vector training accurecy: 0.2195238095238095
KNN  training accurecy: 0.4202292768959436
decision tree accurecy: 0.9723280423280424
logistic regression accurecy: 0.1428395061728395
random forest accurecy: 0.9723104056437389
naive bayes accurecy: 0.11627865961199295
GradientBoostingClassifier  training accurecy: 0.26763668430335097
```

Fig 5.2.1 :All classifier training Accuracy.

Table 1 : Training/validation accuracy and loss



Training loss	Training accuracy	Validation loss	Validation accuracy
0.5900	0.6845	0.4306	0.66
0.4352	0.8182	0.3484	0.65
0.3475	0.8706	0.2915	0.66
0.2893	0.8982	0.2762	0.654
0.2553	0.9128	0.2801	0.66

Random forest gives 97.23 percent and decision tree gives 97.23 percent. But at this time another classifier is not predicting very good performance which is shown in fig 5.2.1.

Table 1 : Discussion based on classification report .

Algorithm	Result	Discussion
SVM	21.95 %	Performance Bad
Decision Tree	97.23 %	Performance Good
Random Forest	97.23 %	Performance Good
KNN	42.02 %	Performance Not very good
Gradient boosting	26.76 %	Performance Bad

Here we try to Discuss which algorithm gives us better performance. Its means is not that other algorithms are not bad. Algorithms are given up better performance when we do the right preprocessing for a specific algorithm like some algorithms perform well if all values are numeric then they perform very well.so I think it depends on the data set and our preprocessing.

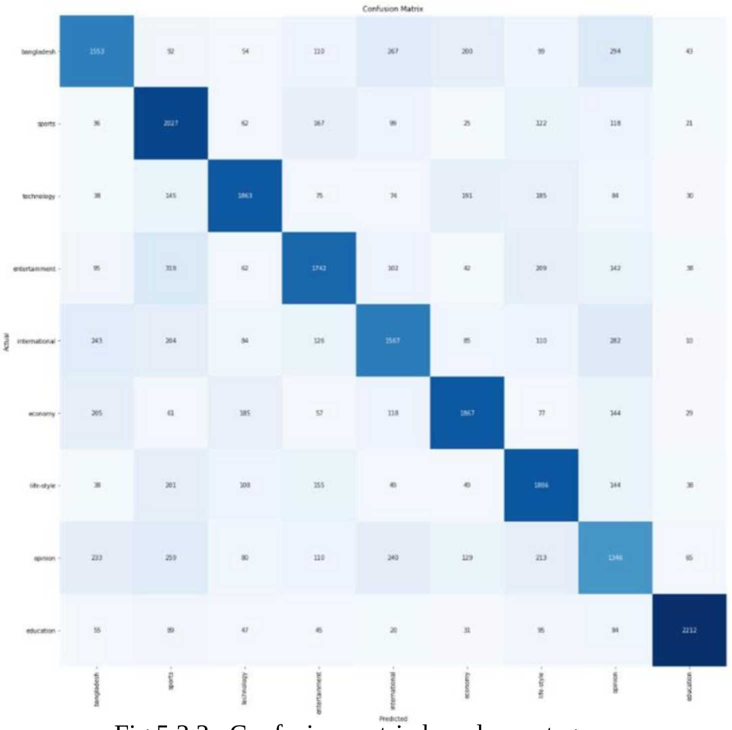
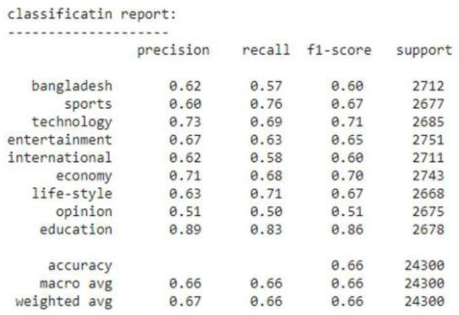


Fig 5.2.2 : Confusion matrix based on category.

Here we see that we have 9 categories and in Confusion matrix helps us to clearly describes which categories can be categorized better as shown in fig 5.2.2. We know that when we give to a category all categories are not accurate then the Confusion matrix helps us to understand easily.



```
classificatin report:
--------------------
               precision    recall  f1-score   support

   bangladesh       0.62      0.57      0.60      2712
       sports       0.60      0.76      0.67      2677
   technology       0.73      0.69      0.71      2685
entertainment       0.67      0.63      0.65      2751
international       0.62      0.58      0.60      2711
      economy       0.71      0.68      0.70      2743
   life-style       0.63      0.71      0.67      2668
      opinion       0.51      0.50      0.51      2675
    education       0.89      0.83      0.86      2678

     accuracy                           0.66     24300
    macro avg       0.66      0.66      0.66     24300
 weighted avg       0.67      0.66      0.66     24300
```

Fig 5.2.3 :Classification report of confusion matrix.

Here, we see that all result precision, recall, f1-score, and support for all specific categories. In our experiment education, economy, and technology categories result are better than other categories that are in Fig 5.2.3.

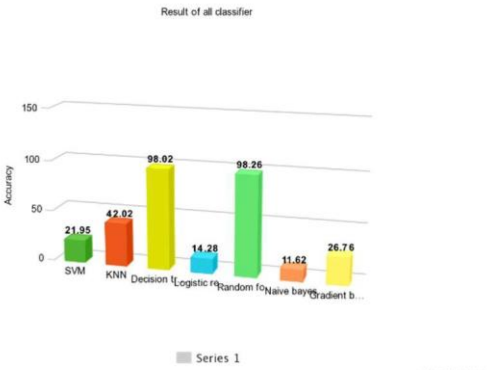


Fig 5.2.4 : Algorithm result in 3d format.

Random forest gives 97.23 percent and decision tree gives 97.23 percent. But at this time another classifier is not predicting very good performance. All results are shown in Fig-5.2.4. In the confusion matrix test accuracy is 66.27 percent.

Discussion

Table 5.2.4 : Experimental result discussion.

Experiment Name	Result Discussion	Comment
Experiment with Random Forest	Random forest avoids overfitting for this reason it works better on our dataset. Its test accuracy is 98%.	In our research, we use some classifiers like SVM, decision tree, random forest, KNN, Gradient boosting, at this time random forest, and decision tree and give the best accuracy in all classifiers. Random forest gives 97.23 percent and decision tree gives 97.23 percent. But this time SVM 21.2%, Logistics regression 14.7%, Naive Bayes 11.6%, and Gradient boosting 21.76% testing accuracy that is shown in fig 5.2.4. As random forest gives us the best output so we pick it for our research.
Experiment with Decision Tree	The Decision tree explains all possible outputs for a single input. So, in our data set it works better. Its test accuracy is 97%.	
Experiment with KNN	KNN visit all nearest neighbor. Its performance is not so good. Its test accuracy is 42.02%.	

CHAPTER 6

Impact on Society and Sustainability

6.1 Impact on Society

Newspapers have become an integral part of our society and be it online or offline platforms. Selecting the right news is a very challenging task. On the other hand, if the news is categorized according to the headline then many people will be interested in reading the news without any kind of boring.

6.2 Ethical aspects

The data we collect for our research purposes will not be used for any unethical purpose. In addition, by categorizing news, people's interest in reading news will increase, and through that everyone will get updated news of the world. This will increase people's desire to write columns and blogs in newspapers. Besides, reading newspapers will enrich the knowledge base.

6.3 Sustainability Plan

In this paper, we will try to work on Bangla news and social media news categorization by using deep learning. We will categorize all news by their headline and we get which category of news it will be. Our target will be to collect more data and process them and get accurate results by applying some models.

CHAPTER 7

Summary, Conclusion, Recommendation, And Implementation for Future Work

7.1 Summary of the study

There are many document categorization systems developed for English language processing but no usable system is developed for Bangla texts. There are many social media or TV channels to publish news but there is a major problem for all people they can’t read previous news when they want. Besides, everyone doesn’t like every category, and some people like some categories. As a result, many people are disturbed when they read newspaper front pages. Every category of news is available there but creates a problem when a user chooses a random category. That’s why we add many types of categories to our project.

7.2 Conclusion

In this research work, we tried to categorize the news headline. That’s why we collected a lot of newspaper headlines from various news portals. The headline input we give for our category prediction is 50-100 words which can be predicted best. Our main focus is to categorize the newspaper article and increase people's desire to read the newspaper.

7.3 Implication for further study

In this work, we collect data only from the Prothom Alo newspaper. In the future, we try to add more Natural language terms. We will work on more Attributes and add many data samples. We also want to work with fake in news detection the future from social media (covid 19)[17].

Novel contribution

In our paper we use Machine Learning and Deep learning both algorithm that are unique from another paper. Our result accuracy is very good than another paper. Our Data amount is 4 lakhs but our accuracy not bad because here we use Tokenizer but another paper they use TF-IDF. Here we try to compare the result of ML and DL. Maximum paper is made by English Data set but we use Bangla Dataset. Our All Data is unique.

Reference:

[1] P. Chowdhury, E. M. Eumi, O. Sarkar, M. Ahamed, et al., “Bangla news classification using glove vectorization, lstm, and cnn,” in Proceedings of the International Conference on Big Data, IoT, and Machine Learning, pp. 723–731, Springer, 2022.

[2] M. Ahmad, F. N. Mishu, and S. Limon, “Bangla news classification using machine learning,” 2018.

[3] M. H. I. Bijoy, M. Hasan, A. N. Tusher, M. M. Rahman, M. J. Mia, and M. Rabbani, “An automated approach for bangla sentence classification using supervised algorithms,” in 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1–6, IEEE, 2021.

[4] M. M. Islam, A. K. M. Masum, M. G. Rabbani, R. Zannat, and M. Rahman, “Performance measurement of multiple supervised learning algorithms for bengali news headline sentiment classification,” in 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), pp. 235–239, IEEE, 2019.

[5] D. Bhakta, A. A. Dash, M. Bari, S. Shatabda, et al., “Supervised machine learning for multi-label classification of bangla articles,” in International Fig. 6. confusion matrix Conference on Cyber Security and Computer Science, pp. 477–487, Springer, 2020.

[6] M. R. Hossain and M. M. Hoque, “Automatic bengali document categorization based on word embedding and statistical learning approaches,” in 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), pp. 1–6, IEEE, 2018.

[7] M. T. Alam and M. M. Islam, “Bard: Bangla article classification using a new comprehensive dataset,” in 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), pp. 1–5, IEEE, 2018.

[8] S. D. Mahajan and D. Ingle, “News classification using machine learning,”

[9] R. A. Naqvi, M. A. Khan, N. Malik, S. Saqib, T. Alyas, and D. Hussain, “Roman urdu news headline classification empowered with machine learning,” Computers, Materials & Continua, vol. 65, no. 2, pp. 1221– 1236, 2020.

[10] M. S. Salayhin, “Development of a bangla news classification system,” 2019.

[11] M. Mansur, Analysis of n-gram based text categorization for bangla in a newspaper corpus. PhD thesis, BRAC University, 2006.

[12] G. Kaur and K. Bajaj, “News classification and its techniques: a review,” IOSR Journal of Computer Engineering (IOSR-JCE), vol. 18, no. 1, pp. 22–26, 2016.

[13] K. Salehin, F. Ahmed, M. Nabi, M. K. Alam, et al., Bangla text classification using machine learning and deep learning techniques. PhD thesis, Brac University, 2021.

[14] A. A. Imran, Z. Wahid, and T. Ahmed, “Bnnet: A deep neural network for the identification of satire and fake bangla news,” in International conference on computational data and social networks, pp. 464–475, Springer, 2020.

[15] K. Salehin, M. K. Alam, M. A. Nabi, F. Ahmed, and F. B. Ashraf, “A comparative study of different text classification approaches for bangla news classification,” in 2021 24th International Conference on Computer and Information Technology (ICCIT), pp. 1– 6, IEEE, 2021.

[16] R. Amin, N. S. Sworna, and N. Hossain, “Multiclass classification for bangla news tags with parallel cnn using word level data augmentation,” in 2020 IEEE Region 10 Symposium (TENSYMP), pp. 174–177, IEEE, 2020.

[17] M. G. Hussain, M. R. Hasan, M. Rahman, J. Protim, and S. A. Hasan, “Detection of bangla fake news using mnb and svm classifier,” arXiv preprint arXiv:2005.14627, 2020.

[18] S. Yeasmin, R. Kuri, A. M. H. Rana, A. Uddin, A. S. U. Pathan, and H. Riaz, “Multi- category bangla news classification using machine learning classifiers and multi-layer dense neural network,”

[19] M. Kowsher, A. Tahabilder, N. Jahan Prottasha, M. Abdur-Rakib, M. Uddin, P. Saha, et al., “Bangla topic classification using supervised learning,” in Computational Intelligence in Pattern Recognition, pp. 505–518, Springer, 2022.

[20] R. Rahman, “A benchmark study on machine learning methods using several feature extraction techniques for news genre detection from bangla news articles & titles,” in 7th International Conference on Networking, Systems and Security, pp. 25–35, 2020.

[21] S. Rahman and P. Chakraborty, “Bangla document classification using deep recurrent neural network with bilstm,” in Proceedings of international conference on machine intelligence and data science applications, pp. 507–519, Springer, 2021.

[22] T. Islam, A. I. Prince, M. M. Z. Khan, M. I. Jabiullah, and M. T. Habib, “An in-depth exploration of bangla blog post classification,” Bulletin of Electrical Engineering and Informatics, vol. 10, no. 2, pp. 742–749, 2021.

[23] A. N. Chy, M. H. Seddiqui, and S. Das, “Bangla news classification using naive bayes classifier,” in 16th Int’l Conf. Computer and Information Technology, pp. 366–371, IEEE, 2014.

[24] A. Ahmad and M. R. Amin, “Bengali word embeddings and it’s application in solving document classification problem,” in 2016 19th International Conference on Computer and Information Technology (ICCIT), pp. 425–430, IEEE, 2016.

[25] P. K. Mallick, S. Mishra, and G.-S. Chae, “Digital media news categorization using bernoulli document model for web content convergence,” Personal and Ubiquitous Computing, pp. 1–16, 2020

[26] Ansary, Adil, and Mohammad Rakib Hassan. "A Comparative Study on Detecting Bangla Fake News on Social Media Using Machine Learning Algorithms." (2021).

[27] KL, Namita Arun Amdalli1 Santhosh Kumar, and Jharna Majumdar. "Classification and Analysis of Online News Articles using NDTV."

[28] Islam, Md Majedul, et al. "Performance Measurement of Multiple Supervised Learning Algorithms for Bengali News Headline Sentiment Classification." *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*. IEEE, 2019.

[29] Biradar, Sampada, and M. M. Raikar. "Performance Analysis of Text Classifiers based on News Articles-A Survey." *Indian Journal of Scientific Research* (2017): 156-162.

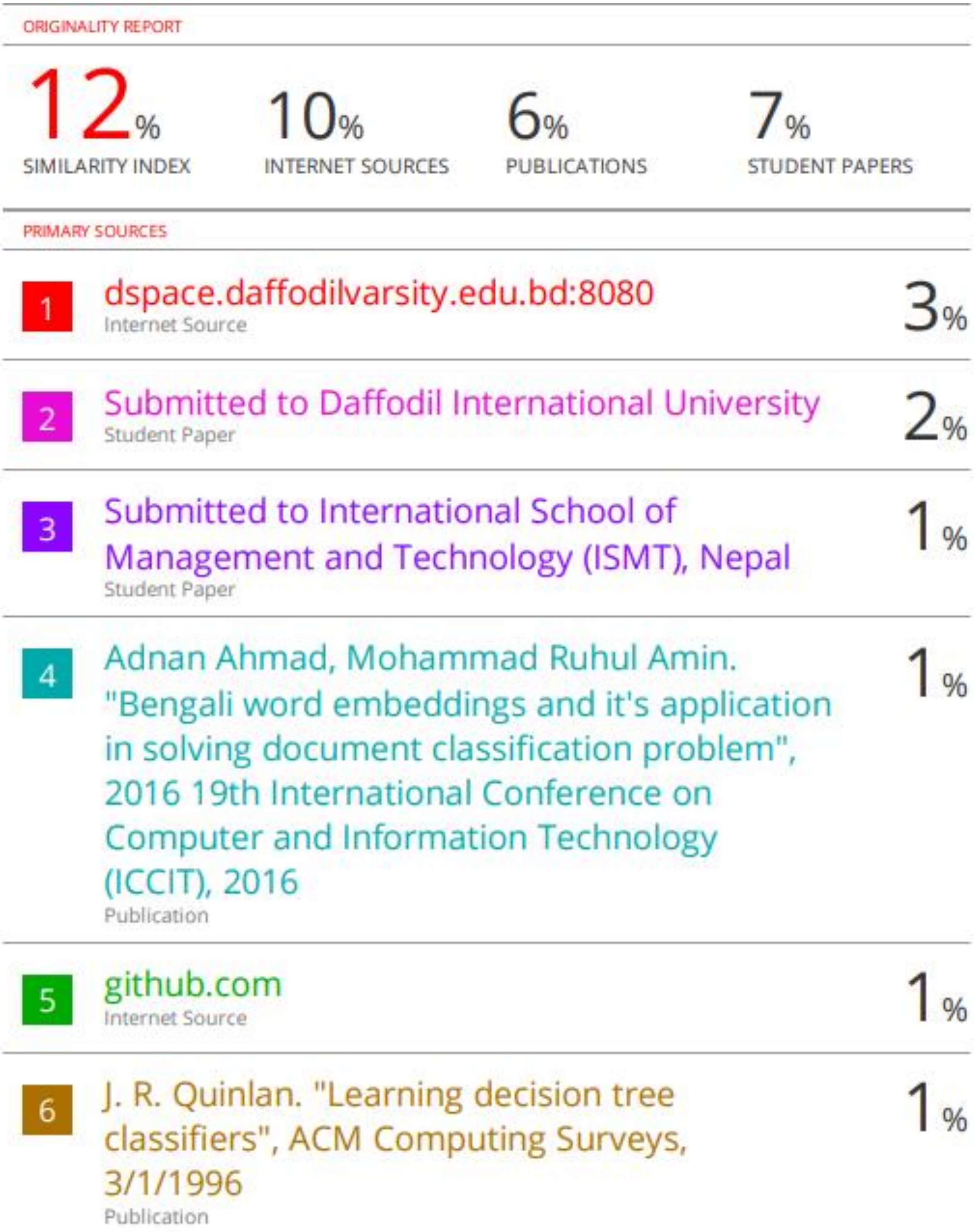
[30] Fauzi, M. Ali, et al. "Indonesian news classification using Naïve Bayes and two-phase feature selection model." *Indonesian Journal of Electrical Engineering and Computer Science* 2.3 (2016): 401-408.

[31] Islam, Farzana, et al. "Bengali fake news detection." *2020 IEEE 10th International Conference on Intelligent Systems (IS)*. IEEE, 2020.

[32] Khan, Md, and Md Rahman. "Bangla Fake News Detection: Machine Learning Perspective." (2021).

[33] Sraboni, Tasnuba, et al. *FakeDetect: Bangla fake news detection model based on different machine learning classifiers*. Diss. Brac University, 2021.

[34] Hossain, Md Rajib, and Mohammed Moshiul Hoque. "Automatic Bengali document categorization based on word embedding and statistical learning approaches." *2018 International*

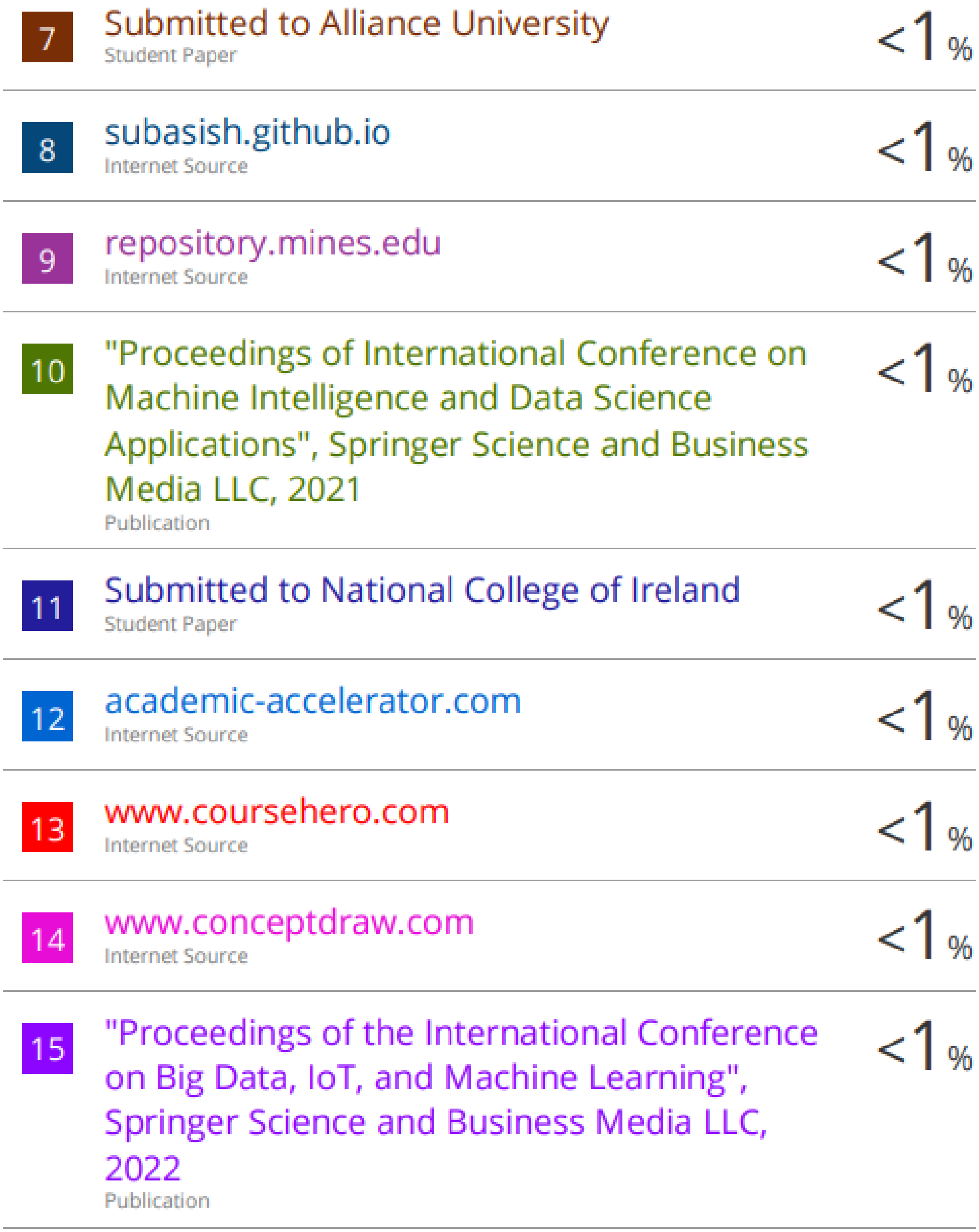


ORIGINALITY REPORT

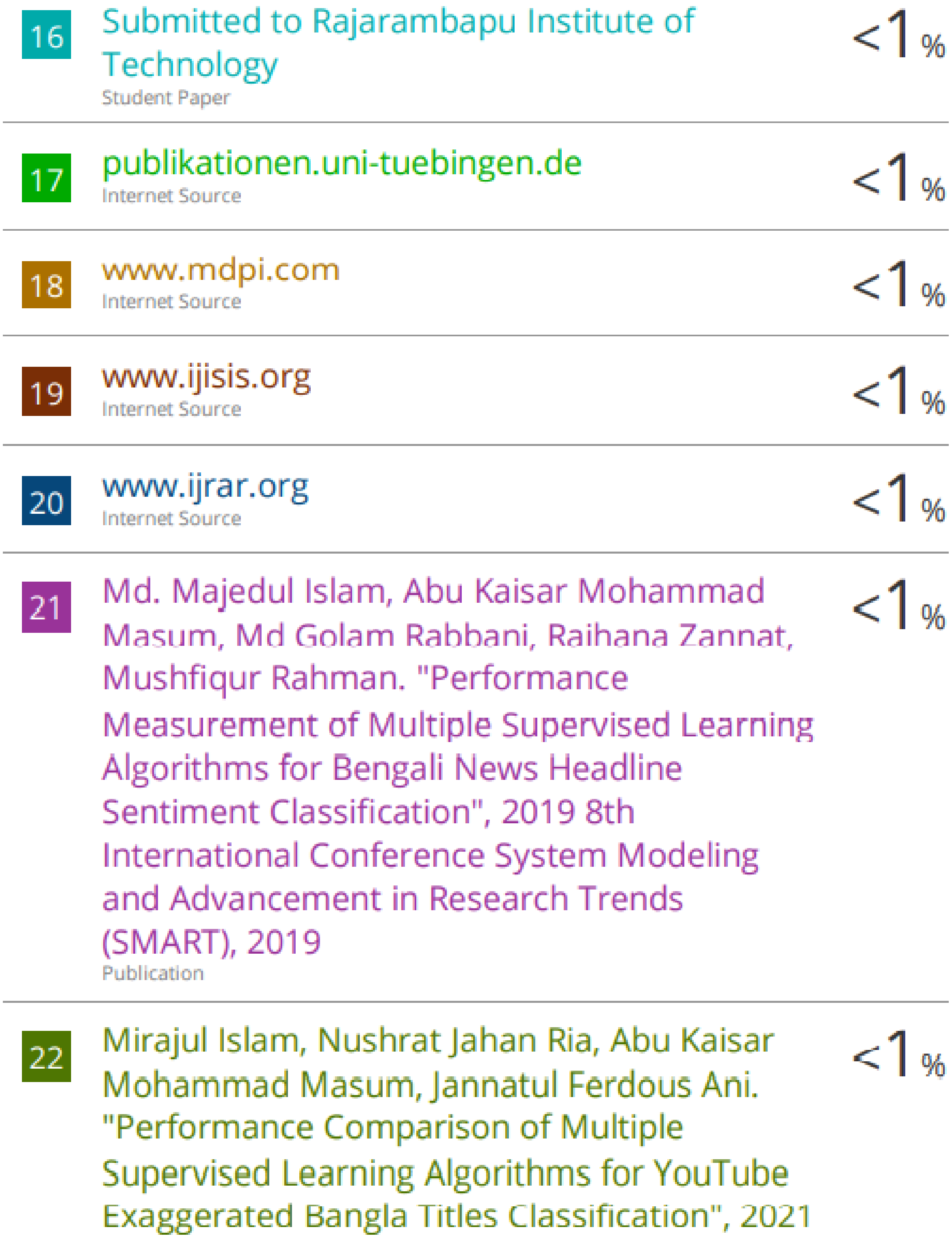
12%	10%	6%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

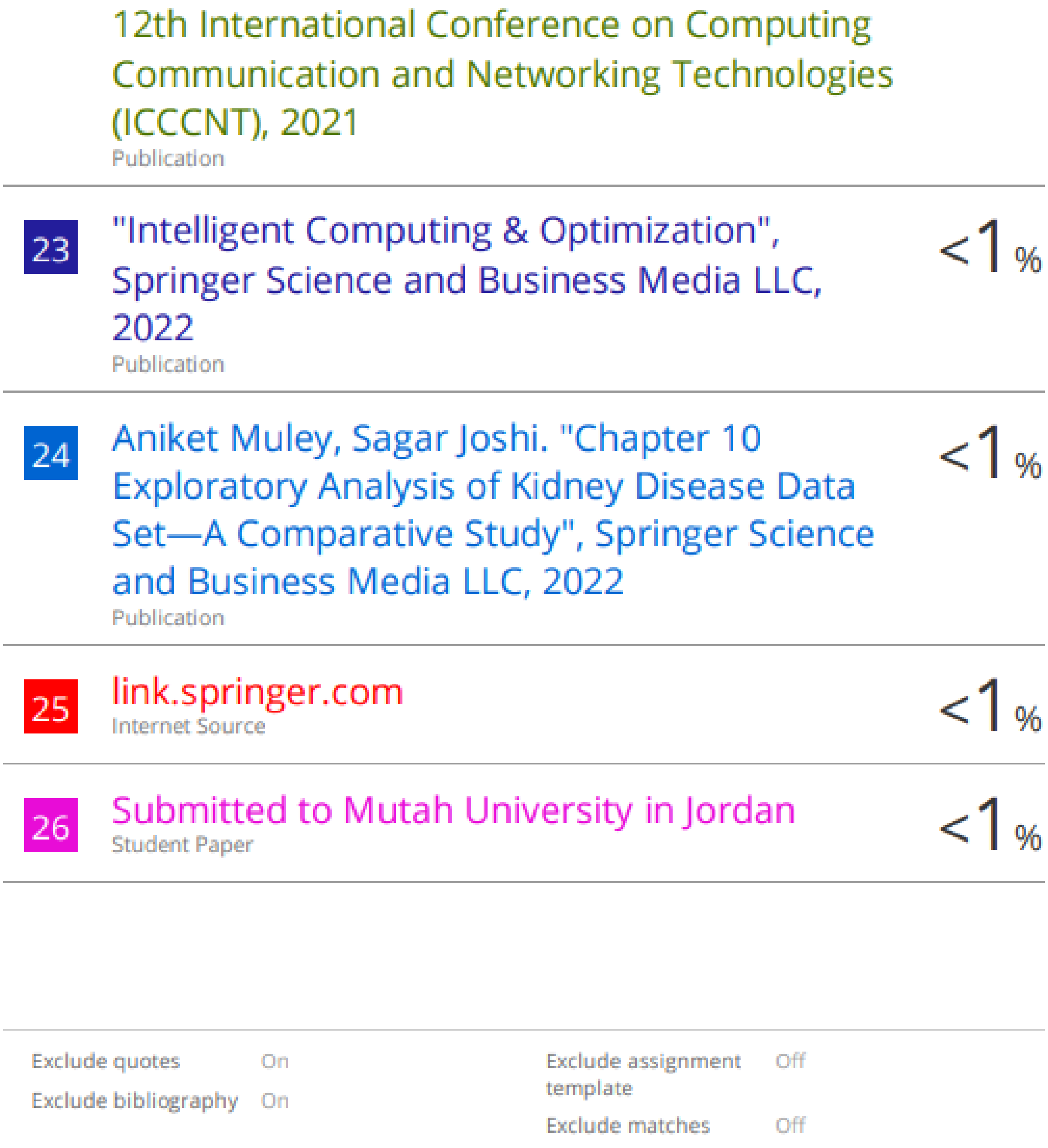
1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
2	Submitted to Daffodil International University Student Paper	2%
3	Submitted to International School of Management and Technology (ISMT), Nepal Student Paper	1%
4	Adnan Ahmad, Mohammad Ruhul Amin. "Bengali word embeddings and it's application in solving document classification problem", 2016 19th International Conference on Computer and Information Technology (ICCIT), 2016 Publication	1%
5	github.com Internet Source	1%
6	J. R. Quinlan. "Learning decision tree classifiers", ACM Computing Surveys, 3/1/1996 Publication	1%



7	Submitted to Alliance University Student Paper	<1 %
8	subasish.github.io Internet Source	<1 %
9	repository.mines.edu Internet Source	<1 %
10	"Proceedings of International Conference on Machine Intelligence and Data Science Applications", Springer Science and Business Media LLC, 2021 Publication	<1 %
11	Submitted to National College of Ireland Student Paper	<1 %
12	academic-accelerator.com Internet Source	<1 %
13	www.coursehero.com Internet Source	<1 %
14	www.conceptdraw.com Internet Source	<1 %
15	"Proceedings of the International Conference on Big Data, IoT, and Machine Learning", Springer Science and Business Media LLC, 2022 Publication	<1 %



16	Submitted to Rajarambapu Institute of Technology Student Paper	<1 %
17	publikationen.uni-tuebingen.de Internet Source	<1 %
18	www.mdpi.com Internet Source	<1 %
19	www.ijisis.org Internet Source	<1 %
20	www.ijrar.org Internet Source	<1 %
21	Md. Majedul Islam, Abu Kaisar Mohammad Masum, Md Golam Rabbani, Raihana Zannat, Mushfiqur Rahman. "Performance Measurement of Multiple Supervised Learning Algorithms for Bengali News Headline Sentiment Classification", 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), 2019 Publication	<1 %
22	Mirajul Islam, Nushrat Jahan Ria, Abu Kaisar Mohammad Masum, Jannatul Ferdous Ani. "Performance Comparison of Multiple Supervised Learning Algorithms for YouTube Exaggerated Bangla Titles Classification", 2021	<1 %



	12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021 Publication	
23	"Intelligent Computing & Optimization", Springer Science and Business Media LLC, 2022 Publication	<1 %
24	Aniket Muley, Sagar Joshi. "Chapter 10 Exploratory Analysis of Kidney Disease Data Set—A Comparative Study", Springer Science and Business Media LLC, 2022 Publication	<1 %
25	link.springer.com Internet Source	<1 %
26	Submitted to Mutah University in Jordan Student Paper	<1 %

Exclude quotes	On	Exclude assignment template	Off
Exclude bibliography	On	Exclude matches	Off