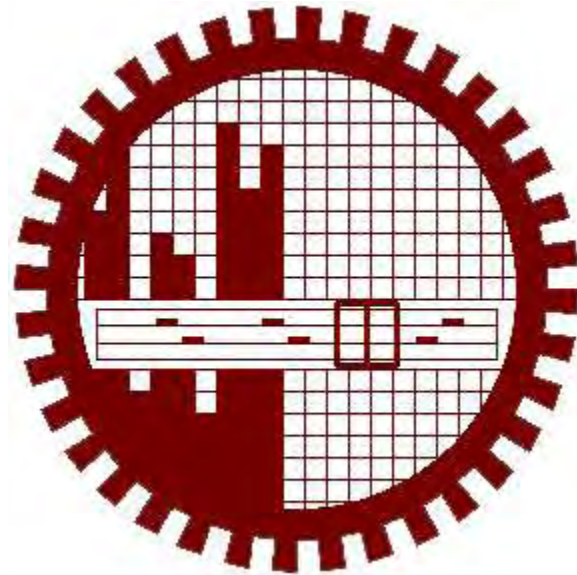


DEVELOPMENT OF A BANGLA NEWS CLASSIFICATION SYSTEM

By

Md Sirajus Salayhin

**POST GRADUATE DIPLOMA IN INFORMATION AND COMMUNICATION
TECHNOLOGY**




Institute of Information and Communication Technology

BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY


March, 2019

The project titled “DEVELOPMENT OF A BANGLA NEWS CLASSIFICATION SYSTEM” submitted by Student Md Sirajus Salayhin, Roll No: 1014311039, Session October 2014, has been accepted as satisfactory in partial fulfillment of the requirements for the degree of Post Graduate Diploma in Information and Communication Technology on 31 March, 2019.


BOARD OF EXAMINERS

1. 

Dr. Md. Saiful Islam **Chairman**
Professor
IICT, BUET, Dhaka
(Supervisor)

2. 

Dr. Shahin Akhter **Member**
Assistant Professor
IICT, BUET, Dhaka

3. 

Dr. Mohammad Arifuzzaman **Member**
Lecturer
IICT, BUET, Dhaka

CANDIDATE'S DECLARATION

It is declared that project or any part of it has not been submitted elsewhere for the award of any degree or diploma.

Signature

S. Sirajus Salayhin

Md Sirajus Salayhin

Table of Contents

Chapter 01	2
Introduction	2
1.1 Background	2
1.2 Present state of the problem	3
1.3 Objective with specific aims and possible outcome	4
Chapter 2	6
Overview of Classification Algorithms	6
2.1 Rule-based Systems	6
2.2 Machine Learning Based Systems	7
2.4 Text Classification Algorithms	9
2.4.2 Support Vector Machines	9
2.4.3 Decision Tree	10
Chapter 3	14
Research Methodology	14
3.2 Preparation of Data	15
3.2.1 Collection of Bangla News	15
3.2.2 Remove Stop Words & Punctuation	16
3.2.3 Stemming	16
3.2.4 Eliminating Irrelevant Words	16
3.3 Feature Extraction	17
3.3.1 Calculating tf-idf	17
3.4 Building Model	18
3.5 Classifier Fitting	18
3.6 Predict the Category	19
3.7 Results	19
3.7 Summary	19
Chapter 4	21
Experimental Results and Discussion	21

4.2 Cleaning Raw Data	23
4.4 Feature Selection and Extraction	24
4.5 Building Model and Fit Dataset for Classifier	24
4.6 Accuracy of Model	25
4.6.3 Random Forest.....	28
4.6.4 Naive Bayes Classifier.....	30
4.8 Compare Recall	32
4.9 F1 Score for all algorithms	33
4.10 Compare Algorithm Accuracy	33
4.10 Summary.....	33
Chapter 5	34
Conclusion and Future Direction	34
5.1 Introduction	34
5.2 Conclusion	35
5.3 Recommendations	35
5.4 Future Direction.....	35
References	36

List of Tables and Figures

Figure 1: Feature extraction Model -----	15
Figure 2: Feature extraction & classification -----	16
Table 1: Observations of the last ten days -----	19
Figure 3: A decision tree for the concept Play Badminton -----	20
Figure 4: illustrates a learned decision tree -----	21
Figure 5: System Diagram of Our Methodology -----	25
Figure 6: Scrapy Code Snippet -----	30
Figure 7: Sample dataset -----	31
Figure 8: News Category -----	32
Figure 9: Stop Words removal -----	34
Figure 10: Confusion Matrix of Decision Tree -----	35
Figure 11: Classification Report for Decision Tree -----	36
Figure 12: Confusion Matrix for Random Forest -----	37
Figure 13: Classification Report for Random Forest -----	38
Figure 14: Confusion Matrix for Naive Bayes -----	39
Figure 15: Classification report for Naive Bayes -----	39
Table 2: Compare Precision -----	40
Table 3: Compare Recall -----	40
Table 4: F1 Score for all algorithm -----	40
Table 5: Compare Algorithm Accuracy -----	41

List of Abbreviations and Technical Terms

Precision of classifier:

Precision is the fraction of relevant instances among the retrieved instances. Precision is used as a measurement of the relevance.

$$\text{Precision} = (\text{True Positive}) / (\text{True Positive} + \text{False Positive})$$

Recall of classifier:

Recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Recall is used as a measurement of the relevance.

$$\text{Precision} = (\text{True Positive}) / (\text{True Positive} + \text{False Negative})$$

F1 Score: In statistical analysis of binary classification, the F_1 score (also F-score or F-measure)

is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score.

$$F1 = 2 \times [(\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})]$$

Support:

The support is the number of samples of the true response that lie in that class.

Acknowledgement

At first I would like to convey my gratitude to Almighty ALLAH for giving me the opportunity to accomplish this project. I would like to thank my supervisor Dr. Md. Saiful Islam, Professor, Institute of Information and Communication Technology (IICT), Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh. He has assigned me an interesting and useful area, which has an extensive range of relevance in the real world. He has provided all sorts of support regarding the project work. Without his proper guidance, advice, continual encouragement and active involvement in this process of this work, it would have not been feasible.

I am indebted to all the teachers, officers and staff of Information and Communication Technology (IICT) for giving me their kind support and information during my study. I am also grateful to my parents whose continuous support all over my life has brought me this far in my career.

Finally, I tender special thanks to the Almighty that I have been successful in my effort to complete the study.

Abstract

On-line newspapers and digital editions of print newspapers has become more and more popular as technology continues to grow. As the number of popular news articles grows and people also have different interests they want to categorize news to read only their interested topics. Classification of on-line news in the past, has often been done manually. Text classification is a well-studied problem. Several methods have been proposed and many of them can be directly applied to news classification as long as there exists a good set of training documents for each predefined category. To develop a news classifier we can use either Collaborative filtering, Content-based filtering, Subscription-based personalization approach. Among the above three approaches, we have chosen to adopt Content filtering approach to support personalized news classification in Categorizer system. The main difficulty in using this approach is that the training data required to build Bangla text classifiers. Few work has been done with ‘Naive Bayes classifier’, ‘Stochastic Gradient Descent (SGD) classifier’, ‘N-Gram Based Text Categorization’, ‘Support Vector Machine (SVM)’ and ‘K-Nearest Neighbor (KNN)’ etc. for Bangla text classification. We want to Compare different Classifier algorithm to know better classification result and finally build a web application where users can search and read category based Bangla news articles.

In this project an application has been developed to compare different classification algorithms for Bangla news classifier and a web application for categories news. Data collection, analysis and model building part has been developed with Python and Python based machine learning library (Scikit Learn). For development the application I have used Python based web framework. After analysis different algorithms we found Naive Bayes has more accuracy then other machine learning algorithms.

Chapter 01

Introduction

1.1 Background

The Web is quickly moving towards a stage for mass joint effort in content generation and utilization, and the expanding number of individuals are swinging to online hotspot for every day news. Bangla online paper began showing up at about a similar time that the web ended up open in Bangladesh. Consistently, huge measure of Bangla news articles are made by the few news destinations that exist in the World-Wide-Web also, its rate increments exponentially.

An online paper has numerous structures. One shape is electronic version of the printed paper. The shopper can peruse the web version much like a paper release; there's no classification, neither with appreciate to content material nor with respect to design. some other type of online paper is news portals or website, which empowers the individual perusing in menus which may be composed in issue classifications and sub-classes. typical for limit of the above assortments of online papers is that the client is accepted to inspect the content from a PC screen.

As the number of popular news articles grows and people also have different interest they want to categorize news to read only their interested topics. Classification of online news in the past, has often been done manually. Text classification is a well-studied problem. Several methods has been proposed and many of them can be directly applied to news classification as long as there exists a good set of training documents for each predefined category. To develop a news classifier we can use either Collaborative filtering, Content-based filtering, Subscription-based personalization approach. Among the above three approaches, we have chosen to adopt Content filtering approach to support personalized news classification in Categorizer system. The main difficulty in using this approach is that the training data required to build Bangla text classifiers. Few work has been done with „Naive Bayes classifier“, „Stochastic Gradient Descent (SGD), classifier“, „N-Gram Based Text Categorization“, „Support Vector Machine (SVM)“ and „K-Nearest Neighbor (KNN)“ etc. for Bangla text classification. We want to Compare different Classifier algorithm to know better classification result and finally build a web application where user can search and read category based Bangla news articles.

1.2 Present state of the problem

Many research based works have been accomplished for English language. Writing audit demonstrates that many built up directed learning calculation have been utilized for content or report classification. Most much of the time utilized systems are K-Nearest Neighbor (KNN), Naive Bayes (NB), N-Grams, Decision Tree (DT), Neural Network (NNet), Support Vector Machines (SVM) and so forth. NB is the most as often as possible utilized methodologies for content arrangement and it is simple for execution and calculation. In, they have utilized Naive Bayes strategy to sort the substance in a site and they got 80% exactness for ten classes. An extensive number of near investigations are additionally done in English archive classification. For example, the creator in utilized KNN, NB and Term Gram for this assignment. They demonstrated a near report where the precision of KNN is a superior decision than NB and Term Gram.

In some exploration indicates new gatherings dataset they demonstrated SVM performed far superior than the various methodologies they have utilized. In spite of the fact that the greater part of work has been accomplished for English dialects however there are additionally some different works which have been accomplished for some different dialects. For example, for Arabic language, SVM in and NB in have been utilized for programmed content characterization. For Tamil language NNet was utilized in with vector space show. Another work was done in for Punjabi language. Hardly any works have been accomplished for classify Bangla language records classification. Another work which have been finished utilizing N-gram technique to sort Bangla day by day paper Corpus. Additionally in paper creators demonstrated that SVM works superior to anything other administered learning for expansive example set of order and got the calculation from auxiliary hazard minimization hypothesis. In our work, we have utilized some entrenched procedures, for example, Random Forest, SVC, Linear SVC, Naive Bayes. Among all Naive Bayes performed much better than the various procedures with exactness 85%.

1.3 Objective with specific aims and possible outcome

The main aim of this project is to build a model for Bangla news classifier and a web application for categories news. To fulfill the aim, the following objectives are identified

1. To study how to classify or categorize Bangla news using some classifier algorithm.
2. Compare different Classifier algorithm to know better classification result.
3. Build a web application where news will be categorized.
4. Searching Bangla news with specific category from the web application.

1.4 Summary

Filtering approach is one of the best approach for news classification in Categorizer system. We want to compare different classifier algorithm for news classification in categorizer system. And we want to know better classification result and finally build a web application where user can search and read category based Bangla News Articles.

Chapter 2

Overview of Classification Algorithms

Content characterization should be possible in two distinctive ways manual and programmed grouping. In the previous, a human annotator deciphers the substance of content and sorts it in like manner. This strategy more often than not can give quality outcomes yet now is the ideal time expending and costly. The last applies AI, characteristic language preparing, and different procedures to naturally characterize message in a quicker and more financially cost-effective way. There are many approaches to automatic text classification, which can be grouped into three different types of systems

- Rule-based systems
- Machine Learning based systems
- Hybrid systems

2.1 Rule-based Systems

Rule based methodologies order content into sorted out gatherings by utilizing a lot of high quality etymological standards. These principles educate the framework to utilize semantically applicable components of a content to distinguish important classes dependent on its substance. Each standard comprises of a precursor or design and an anticipated classification.

State that we need to characterize news articles into 2 gatherings, specifically, Sports and Politics. To start with, you'll have to characterize two arrangements of words that describe each gathering (for example words identified with games, for example, football, ball, LeBron James, and so forth., and words identified with legislative issues, for example, Donald Trump, Hillary Clinton, Putin, and so on.). Next, when you need to group another approaching content, you'll have to tally the quantity of game related words that show up in the content and do likewise for governmental issues related words. In the event that the quantity of game related word appearances is more prominent than the quantity of governmental issues related word tally, at that point the content is delegated sports and the other way around.

Rule based frameworks are human understandable and can be improved after some time. In any case, this methodology has a few hindrances. First off, these frameworks require profound learning of the area. They are additionally tedious, since creating rules for an unpredictable framework can be very testing and as a rule requires a great deal of examination and testing. Guideline based frameworks are additionally hard to keep up and don't scale well given that including new principles can influence the aftereffects of the previous tenets.

2.2 Machine Learning Based Systems

Rather than depending on physically created principles, content order with AI figures out how to mention arrangements dependent on past objective facts. By utilizing pre-named precedents as preparing information, an AI calculation can gain proficiency with the distinctive relationship between bits of content and that a specific yield (for example labels) is normal for a specific information.

The initial move towards preparing a classifier with AI is include extraction a strategy is utilized to change every content into a numerical portrayal as a vector. A standout amongst the most as often as possible utilized methodologies is pack of words, where a vector speaks to the recurrence of a word in a predefined lexicon of words.

For instance, on the off chance that we have characterized our lexicon to have the accompanying words {This, is, the, not, amazing, awful, basketball}, and we needed to vectorize the content "This is marvelous", we would have the accompanying vector portrayal of that content (1, 1, 0, 0, 1, 0, 0).

At that point, the AI calculation is encouraged with preparing information that comprises of sets of capabilities (vectors for every content model) and labels to create an arrangement display

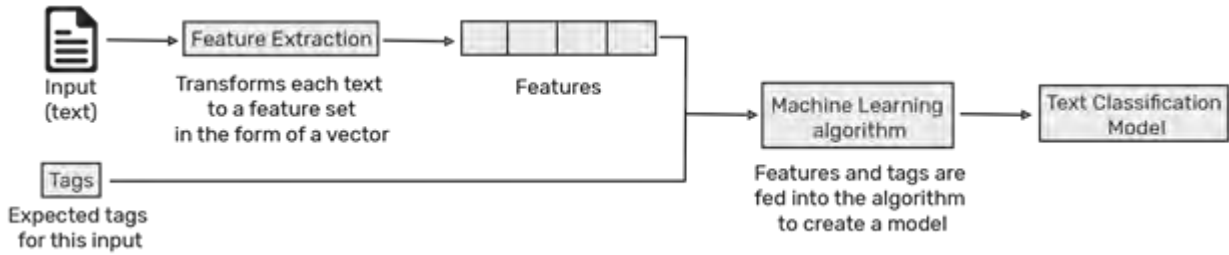


Figure 1 Feature extraction model

When it's prepared with enough preparing tests, the AI model can start to make precise forecasts. A similar element extractor is utilized to change concealed content to include sets which can be bolstered into the characterization model to get expectations on labels (for example sports, legislative issues)



Figure 2 Feature extraction & classification

Text Classification with AI is generally considerably more precise than human-made standard frameworks, particularly on complex grouping undertakings. Additionally, classifiers with AI are simpler to keep up and you can generally label new guides to adapt new assignments.

2.3 Hybrid Systems

Hybrid frameworks join a base classifier prepared with AI and a standard based framework, which is utilized to additionally improve the outcomes. These mixture frameworks can be effectively adjusted by including explicit guidelines for those clashing labels that haven't been accurately displayed by the base classifier.

2.4 Text Classification Algorithms

The absolute most well known AI calculations for making content grouping models incorporate the credulous bayes group of calculations, bolster vector machines, and profound learning.

2.4.1 Naive Bayes

Naive Bayes is a group of measurable calculations we can utilize while doing content characterization. One of the individuals from that family is Multinomial Naive Bayes (MNB). One of its principle points of interest is that you can get great outcomes when information accessible isn't much (~ a few thousand labeled examples) and computational assets are rare.

All you have to know is that Naive Bayes depends on Bayes' Theorem, which encourages us register the contingent probabilities of event of two occasions dependent on the probabilities of event of every individual occasion. This implies any vector that speaks to a content should contain data about the probabilities of appearance of the expressions of the content inside the writings of a given class with the goal that the calculation can register the probability of that content's having a place with the classification.

2.4.2 Support Vector Machines

Support Vector Machines (SVM) is simply one out of many algorithms we can select from when doing textual content classification. Like naive bayes, SVM doesn't need an awful lot training information to begin imparting correct results. Although it needs extra computational sources than Naive Bayes, SVM can attain greater accurate results.

In short, SVM takes care of drawing a "line" or hyperplane that divides a space into two subspaces one subspace that consists of vectors that belong to a crew and every other subspace that carries vectors that do not belong to that group. Those vectors are representations of your coaching texts and a team is a tag you have tagged your texts with.

2.4.3 Decision Tree

A decision tree is a tree-like plan with nodes representing the area the place we pick out an attribute and ask a question; edges represent the answers to the question; and the leaves signify the true output or class label. They are used in non-linear choice making with simple linear selection surface. Decision trees classify the examples through sorting them down the tree from the root to some leaf node, with the leaf node imparting the classification to the example. Each node in the tree acts as a test case for some attribute, and each side descending from that node corresponds to one of the feasible solutions to the test case. This system is recursive in nature and is repeated for each subtree rooted at the new nodes.

Let's illustrate this with assist of an example. Let's anticipate we choose to play badminton on a specific day — say Saturday — how will you determine whether or not to play or not. Let's say you go out and test if it's warm or cold, test the velocity of the wind and humidity, how the climate is, i.e. is it sunny, cloudy, or rainy. You take all these factors into account to figure out if you favor to play or not. So, you calculate all these elements for the final ten days and structure a search for desk like the one below

Day	Weather	Temperature	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No

Table 1. Observations of the last ten days.

Now, we may additionally use this table to figure out whether to play or not. But, what if the climate pattern on Saturday does no longer in shape with any of rows in the table? This can also be a problem. A choice tree would be a brilliant way to represent statistics like this because it takes into account all the feasible paths that can lead to the remaining choice by following a tree-like structure.

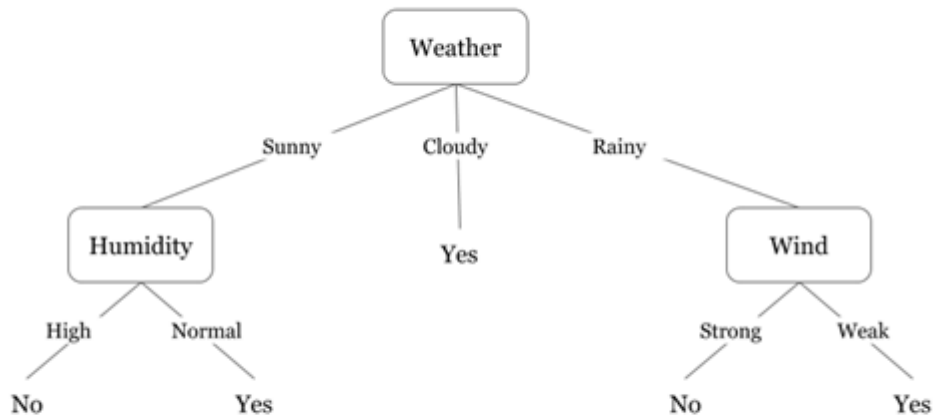


Figure 3 A decision tree for the concept Play Badminton

We can see that every node represents an attribute or feature and the department from every node represents the outcome of that node. Finally, its the leaves of the tree the place the final choice is made. If aspects are continuous, interior nodes can take a look at the fee of a function towards a threshold (Fig. 2). Fig. A selection tree for the idea Play Badminton (when attributes are continuous)

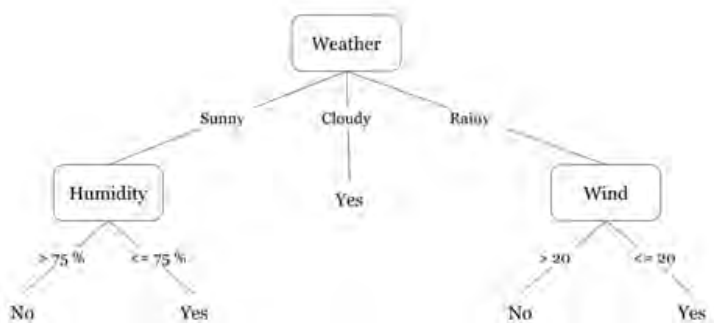


Figure 4 illustrates a learned decision tree.

A general algorithm for a decision tree can be described as follows

1. Pick the best attribute/feature. The best attribute is one which best splits or separates the data.
2. Ask the relevant question.
3. Follow the answer path.
4. Go to step 1 until you arrive to the answer.

The best split is one which separates two different labels into two sets.

2.4.4 Random Forest

Random forest algorithm has grown to be the most frequent algorithm to be used in ML competitions like Kaggle competitions. If you ever search for an easy to use and correct ML algorithm, you will without a doubt get random woodland in the pinnacle results. To apprehend Random forest area algorithm you have to be acquainted with choice trees at first.

Random forest is a tree-based algorithm which entails building countless bushes (decision trees), then combining their output to enhance generalization ability of the model. The technique of combining bushes is acknowledged as an ensemble method. Ensembling is nothing but a combination of weak learners (individual trees) to produce a sturdy learner. Say, you choose to watch a movie. But you are uncertain of its reviews. You ask 10 human beings who have watched the movie. eight of them stated " the movie is fantastic." Since the majority is in favor, you figure out to watch the movie. This is how we use ensemble techniques in our everyday existence too. Random Forest can be used to solve regression and classification problems. In regression problems, the structured variable is continuous. In classification problems, the based variable is categorical.

Summary

In this chapter we have discussed about different classification algorithms and how those algorithms works with couple of examples.

Chapter 3

Research Methodology

3.1 Our Methods

In this segment, we give the details of our approach. Fig. demonstrates the means of the technique followed. In the first place, we utilized an apparatus for gathering URLs of news articles from well-known news gives in Bangla and put away Bangla articles in the database. At that point, we performed pre-preparing of the dataset and performed highlight extraction on the dataset. From that point onward, we connected element choice on the dataset to locate the ideal arrangement of highlights. A few classifiers were then tried and the best model was put away. At last, in light of the best model we built up the apparatus and tried our model that utilized similar highlights.

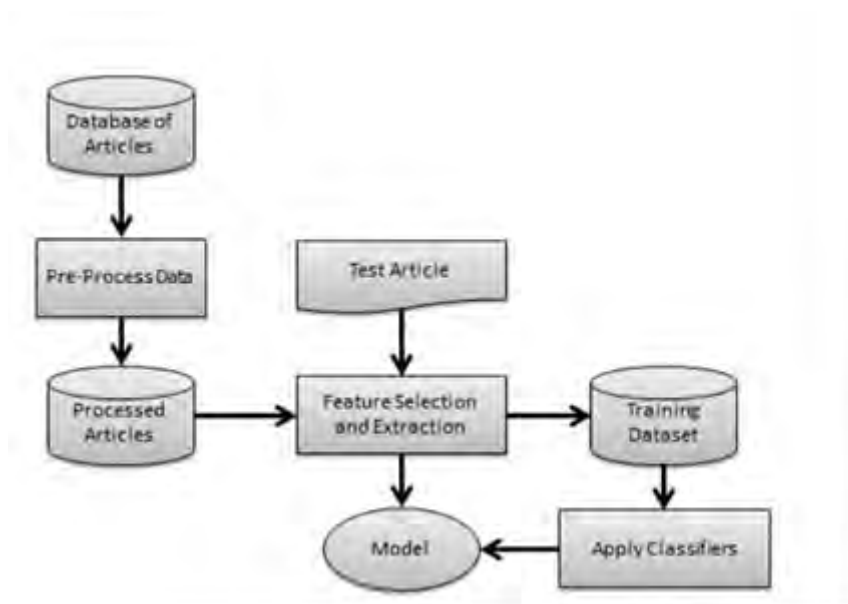


Figure 5 System Diagram of Our Methodology

We have used Python libraries for data collection and model building.

3.2 Preparation of Data

Data Collection is a huge and difficult process. To collect the right amount of data and process it into a right form is a big task. Our data processing had several steps as describe in this section.

3.2.1 Collection of Bangla News

Our first task was to collect articles from online news provider prothom-alo, bdnews24, amader shomoy etc. To do that, we have built a scraper to collect data from those sites and stored as csv file.

3.2.2 Remove Stop Words & Punctuation

After collecting articles our first task is to remove stop words & punctuation mark from news content. And meaningless symbols and Punctuations like [. , ” „ | “ { } () !] have been removed to make the dataset noise free. In this task each and every symbol was replaced by a space to split the sentence into words.

3.2.3 Stemming

After removing punctuations we extracted the unigrams from articles separated by a white space and stored them. For each unigrams those we kept we looked for another unigram within each article and if it contained unigram then replaced with the token.

The algorithm works as below

For each article

For each unigram

If article contains unigram+bivokti

Then replace with unigram

End for

End for

The process is known as finding root words, So that we can reduce the dimensions of final dataset.

3.2.4 Eliminating Irrelevant Words

After finding the base words we created a list of unnecessary words which are not important or irrelevant to categorize text document. The list includes numbers, pronouns, conjunctions and some other single letter words.

3.3 Feature Extraction

The n-grams normally are gathered from a content or discourse corpus. At the point when the things are words, n-grams may likewise be called shingles. Utilizing Latin numerical prefixes, a n-gram of size 1 is alluded to as a 'unigram'; measure 2 is a "bigram"; estimate 3 is a "trigram". Here we create unigram show by part each word with a space and make a bordering succession of n things from a given arrangement of content records. All the more unequivocally we split every one of the sentences into words in a particular content report by utilizing unigram. The primary reason for utilizing n-gram is to change over the archives into arrangement of words where we can figure the estimation of tf-idf effectively.

3.3.1 Calculating tf-idf

TF-IDF is a data recovery procedure that gauges a terms recurrence (TF) and its opposite archive recurrence (IDF). Each word or term has its individual TF and IDF score. The result of the TF and IDF scores of a term is known as the TF-IDF weight of that term. The higher the TF-IDF score, the rarer the term and the other way around. It is a numerical measurement that is expected to reflect how vital a word is to an archive in a gathering or corpus. Usually utilized as a weighting factor in inquiries of data recovery, content mining, and client displaying. The tf-idf esteem builds relatively to the occasions a word shows up in the archive and is balanced by the recurrence of the word in the corpus, which alters for the way that a few words seem all the more as often as possible when all is said in done. Tf-idf is a standout amongst the most famous term-weighting plans today; 83% of content based recommender frameworks in computerized libraries use tf-idf. On account of the term recurrence $tf(t,d)$, the most straightforward decision

is to utilize the crude include of a term in an archive, the occasions that term t happens in record d .

$$TF(t) = \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Total number of terms in the document}}$$

The reverse report recurrence is a proportion of how much data the word gives, that is, regardless of whether the term is normal or uncommon over all records. It is the logarithmically scaled opposite part of the records that contain the word, acquired by separating the absolute number of archives by the quantity of reports containing the term, and after that taking the logarithm of that remainder.

$$IDF(t) = \log_e \frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}}$$

In our dataset we have determined tf-idf to discover noteworthy words for each report. Five classifications of archive are available and the marks are [Technology, Sports, Entertainment, Economy and International]. In our dataset we speak to add up to number of words as segments and absolute number records as columns. The all out size of the dataset is 40000 lines.

3.4 Building Model

After effectively highlight extraction, we are prepared for structure our model. Also, this is being practiced via preparing our machine. We split our dataset into 3:1. The three part of our informational collection are utilized for our preparation dataset and the rest parcel is for trying. That implies, 75% information from the datasets are utilized preparing and rest 25% is considered as the testing.

3.5 Classifier Fitting

In this stage, our machine is prepared or fit for the classifier. We utilize a few classifier, for example, Naïve Byes, Decision Tree, K-Nearest Neighbors, Support Vector Machine and Random Forest for arrange our news content. Sklearn has worked in classifier of this. We simply import it and fit it.

3.6 Predict the Category

This is the last phase of our news order approach. In this stage, our model is being set up for testing Bangla content info information. As per the given info message, this model can arrange this content utilizing a few classifier, for example, Naïve Byes, Decision Tree, K-Nearest Neighbors, Support Vector Machine and Random Forest.

3.7 Results

A few testing techniques are being utilized in the writing for looking at between grouping calculations. In this paper, we have utilized three methods preparing set approval, rate split approval and cross-overlap approval. In preparing set validation, the entire preparing set is utilized for testing

and in this way have higher odds of over-fitting. On account of rate split, the preparation set is isolated into two sections preparing and approval. The model is found out utilizing the preparation set and the approval set is utilized for testing reason. In this paper, we have utilized a 75% preparing and 25% approval for the rate split test. In k-overlay cross approval, the entire dataset is isolated into k break even with formed sub datasets and in every cycle one subset is utilized as the approval set and the rest as preparing set. In this paper, we have utilized a five-overlap cross approval. Since, this is a multi-class order issue, we have utilized the normal of the correctness's accomplished in every classification.

3.7 Summary

In this chapter we have discussed our research methodology and approach about data collection, data cleaning and how we want to implement classification models. Also how we can identify best model for Bangla News Classification.

Chapter 4

Experimental Results and Discussion

4.1 Collecting Dataset

This chapter 4 mainly focuses on the descriptive analysis of the data used in the research as well as the experimental results of our project. Raw data are from the most renowned news portal of Bangladesh named Prothom Alo, Bdnews24, Amader Somoy etc. We collect our data by using Python scrapy framework. After collecting data, news is stored on csv file. In these file, data are present with some html tag name. While collecting those dataset we have label them into their specific category. Here is a snippet of our scraper which is written in Python.

```
import datetime

import scrapy

from corpus_builder.templates.spider import CommonSpider

class ProthomAloSpider(CommonSpider):
    name = "prothom_alo"
    allowed_domains = ["prothomalo.com"]
    base_url = "https://www.prothomalo.com"
    allowed_configurations = [
        ['start_page', 'end_page'],
        ['start_page'],
        ['archive', 'start_date'],
        ['archive', 'start_date', 'end_date'],
        ['category', 'start_page'],
        ['category', 'start_page', 'end_page']
    ]

    start_request_url = base_url

    content_body = {
        'xpath': '//article//p//text()'
    }

    def request_index(self, response):
        if not self.archive:
            categories = response.xpath('//ul[@id=1]/li/a/@href').re('[a-z0-9]*$')
            categories.remove('todays-paper')

            # implement someday
            # subcategories = response.xpath('//ul[@id=1]/li/a[@href="{0}"]/../ul/li/a/@href').re('[a-z0-9]*$.format(
            #     category)

        if self.category:
            if self.category in categories:
                categories = [self.category]
            else:
                raise ValueError('invalid category slug. available slugs: {}'.format(' '.join(categories)))
```

Figure 6 Scrapy Code Snippet

After collecting the dataset we have load the dataset in Jupyter Notebook.

	category	content
39995	economy	হরতাল অবরোধের কারণে ইউরোপের ক্রেতাদের জেট অ্যাকাউন্ট অন ফায়ার অ্যান্ড বিকিং সফটওয়্যার বাংলাদেশের পোশাক কারখানা পরিদর্শন কার্যক্রম বাধাগ্রস্ত হচ্ছে। বর্তমান রাজনৈতিক পরিস্থিতিতে বাংলাদেশের পোশাকশিল্পের সার্বিক অবস্থা নিয়ে প্রতিবেদনে এমন তথ্যই তুলে ধরেছে কানাডার দৈনিক পত্রিকা টরেন্টো স্টার। পত্রিকাটিকে অ্যাকাউন্টের নির্বাহী পরিচালক রব ওয়েস বলেছেন, হরতাল অবরোধ কারখানা পরিদর্শন করাটা চ্যালেঞ্জ হয়ে দাঁড়িয়েছে। কয়েক সপ্তাহ আগে শুরু হলেও এ পর্যন্ত মাত্র ১০টি কারখানার পরিদর্শন কার্যক্রম শেষ হয়েছে। রব ও...
39996	bangladesh	সড়কজুড়ে ছোট বড় অসংখ্য গর্ত। পিচ ঢালাইয়ের অক্টিভিটি নেই। ধূলাবালু উড়ছে সমানে। গাড়ি চলাচল তো দূরে, হেটে চলাচল করাও দুর্কহ হয়ে পড়েছে। এই চিত্র লক্ষ্মীপুরের রামগঞ্জ উপজেলার রামগঞ্জ ওয়াপদা সড়কের। ছয় কিলোমিটারের এই সড়ক দিয়ে যাতায়াত লামচর, কাক্রনপুর, চণ্ডীপুর ও ইছাপুর ইউনিয়নের অর্ধলক্ষাধিক মানুষের। শুধু এটি নয়, উপজেলার অভ্যন্তরীণ ১৫০টি সড়কের এখন করণ হাল। সড়কের দুরবস্থার কারণে ১০টি ইউনিয়নের ৩ লক্ষাধিক মানুষ পাঁচ বছর ধরে ভোগান্তি পোহাচ্ছে। এসব সড়কে যাত্রীবাহী বাস ছাড়া অন্য সব যানবাহন চলাচল করে। ...
39997	opinion	মুজিব 'বাংলা দেশ' বেছে নিয়েছিলেন কিলগোর বাঙালির প্রতি মার্কিন রাষ্ট্রদূত এড্‌জু আই কিলগোরের আবেগ অনুভূতিক আমরা আচার ব্লাড, যাকে বলি বাংলাদেশের মুক্তিসংগ্রামে একমাত্র সক্রিয় আমেরিকান মুক্তিযোদ্ধা, তাঁর সঙ্গে তুলনা করতে পারি। ব্লাডের মতোই বাঙালি ও বঙ্গবন্ধুর প্রতি কিলগোরের অনুরাগ লক্ষ করি। ১৯৬৭-৭০ সালে ঢাকার মার্কিন কনসুলেটে কখনো দ্বিতীয় প্রধান, কখনো ভারপ্রাপ্ত কনসাল জেনারেলের দায়িত্ব পালনকারী কিলগোর পরে নিজের দাবি করেছেন, 'আমি শেখ মুজিবের বন্ধু হয়ে উঠেছিলাম।' মুক্তরাষ্ট্রের আলাব...
39998	bangladesh	১৮ ফেব্রুয়ারি ছিল ইমেরিটাস অধ্যাপক আনিসুজ্জামানের ৮১তম জন্মবার্ষিকী। এ উপলক্ষে আজ শনিবার জাতীয় জাদুঘরের মূল মিলনায়তনে সংবর্ধনা অনুষ্ঠানে শুভেচ্ছা জানাতে আসা অনুজ অনুরাগীরা তাঁকে সাহিত্য এবং শিক্ষার বাতিঘর উল্লেখ করে জাতীয় অধ্যাপক ঘোষণা করার জন্য সরকারের কাছে দাবি জানান। অনুষ্ঠানের প্রধান অতিথি আনিসুজ্জামানের অগ্রজ অর্থমন্ত্রী আবুল মাল আবদুল মুহিত বলেন, 'আনিসুজ্জামান আমাদের সাহিত্যের বাতিঘর, এটা নিয়ে কোনো প্রশ্ন নেই। কিন্তু একই সঙ্গে সে বিবেকেরও বাতিঘর।' তিনি বলেন, 'আনিস আমাদের মেহাৎপদ, জন্...
39999	bangladesh	ঢানা হরতালের কারণে বণ্ডার পাইকারি মোকাম মহাঘান বাজার থেকে ঢাকা চট্টগ্রামসহ সারা দেশে সবজির সরবরাহ বন্ধ রয়েছে। পাইকারি বাজার থেকে সরবরাহ না থাকায় শহরের খুচরা বাজারে সবজি পাচ্ছেন না ক্রেতারা। পণ্যবাহী ট্রাকে হরতাল সমর্থকদের হামলা জঙ্কচুরের কারণে মালিকেরা ট্রাক জাড়া দিতে রাজি না হওয়ায় মোকাম থেকে কোনো সবজি পাঠাতে পারছেন না ব্যবসায়ীরা। স্থানীয়রা জানান, বাজারে ক্রেতা না থাকায় খেত থেকে সবজি তুলে ন্যায্য দাম পাচ্ছেন না চাষিরা। আবার সময়মতো সবজি না তোলার কারণে অনেক খেতে সবজি পচে নষ্ট হয়ে য...

Figure 7 Sample dataset

Following data categories are collected for this project

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f20e6ee1ba8>
```

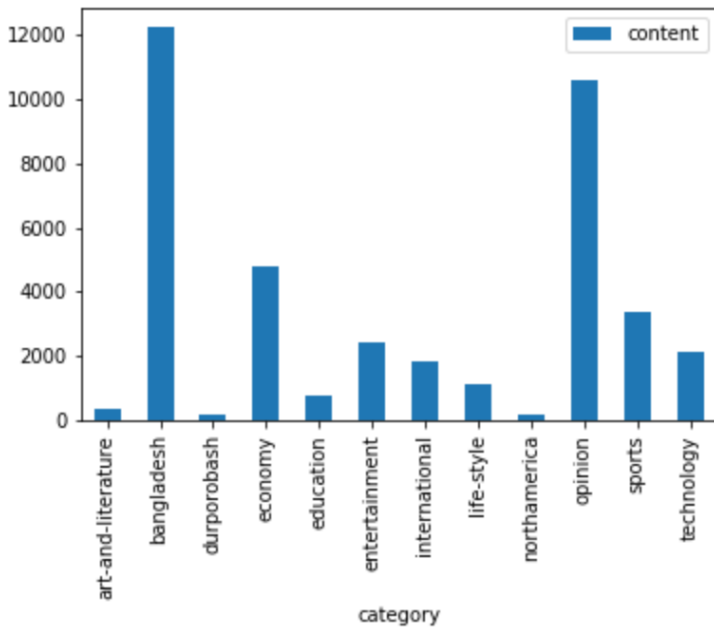


Figure 8 News Category

4.2 Cleaning Raw Data

We use a script file to be helpful of our data pre-processing task.

This python script file is responsible for

- i. Remove all html tag name.
- ii. Remove unnecessary spaces from the text.
- iii. Remove all new line of each news and arrange it in a line
- Iv. While collecting data from website we have assigned article to a specific category

4.3 Stop Words Removal

We develop a python code for classify a news into a category. After joining all news into a file, our system is ready for building a model. For this, a little cleaning process is done before. We create a list that contains some Bangla words that actually no related with the category of the news. We called it as Stop words and named it Excluded words list. Just checking that if stop words are present in our input file or not. If exists, must be removed.

We have found a list of stop words from this git repo

<https://github.com/stopwords-iso/stopwords-bn>

```
# Stop Word List: https://github.com/stopwords-iso/stopwords-bn
import re
stop_words = [
    "অতএব", "অথচ", "অথবা", "অনুযায়ী", "আনেক", "আনোক", "আনোকই", "অস্তুত", "অন্য",
    "অবধি", "অবশ্য", "অর্থাৎ", "আই", "আগামী", "আগে", "আগেই", "আছে", "আজ", "আদ্যভাগে",
    "আপনার", "আপনি", "আবার", "আমরা", "আমাকে", "আমাদের", "আমার", "আমি", "আর",
    "আরও", "ই", "ইত্যাদি", "ইহা", "উচিত", "উত্তর", "উনি", "উপর", "উপরে", "এ", "এদের",
    "এরা", "এই", "একই", "একটি", "একবার", "এক", "একু", "এখন", "এখনও", "এখানে",
    "এখানেই", "এটা", "এটাই", "এটি", "এত", "এতটাই", "এতে", "এদের", "এব", "এবং",
    "এবার", "এমন", "এমনকী", "এমনি", "এর", "এরা", "এল", "এস", "এসে", "ঐ", "ও",
    "ওদের", "ওর", "ওরা", "ওই", "ওকে", "ওখানে", "ওদের", "ওর", "ওরা", "কখনও", "কত",
    "কব", "কমানে", "কয়েক", "কয়েকটি", "করাছে", "করাছেন", "করতে", "করাবে", "করাবেন",
    "করলে", "করালেন", "করা", "করাই", "করায়", "করার", "করি", "করিতে", "করিয়া", "করিয়ে",
    "করে", "করেই", "করেছিলেন", "করেছে", "করেছেন", "করেন", "কড়িকে", "কাছ", "কাছে", "কাজ",
    "কাজে", "কারও", "কারণ", "কি", "কিংবা", "কিছু", "কিছুই", "কিন্তু", "কী", "কে", "কেউ", "কেউই",
    "কোথা", "কেন", "কোটি", "কোন", "কোনও", "কোনো", "ক্ষত্রে", "কয়েক", "খুব", "গিয়ে", "গিয়েছে",
    "গিয়ে", "শুধি", "গোছে", "গেল", "গেলে", "গোটা", "চলে", "চান", "চায়", "চার", "চালু", "চেয়ে",
    "চেষ্টা", "ছাড়া", "ছাড়াও", "ছিল", "ছিলেন", "জন", "জনকে", "জনক", "জনা", "জন্যে", "জানতে",
    "জানা", "জানানো", "জানায়", "জানিয়ে", "জানিয়েছে", "জে", "জনজন", "টি", "ঠিক", "তখন",
    "তত", "তথা", "তবু", "তবে", "তা", "তাকে", "তাদের", "তার", "তার", "তার", "তাঁহারা", "তাই",
    "তাও", "তাকে", "তাতে", "তাদের", "তার", "তারপর", "তার", "তার", "তাহলে", "তাঁহা", "তাহাতে",
    "তাহার", "তিনএ", "তিনি", "তিনিও", "তুমি", "তুলে", "তমনে", "তো", "তোমার", "থাকবে", "থাকবেন",
    "থাকা", "থাকায়", "থাকে", "থাকেন", "থেকে", "থেকেই", "থেকেও", "দিকে", "দিতে", "দিন", "দিয়ে",
    "দিয়েছে", "দিয়েছেন", "দিলেন", "দু", "দুই", "দুটি", "দুটাই", "দেওয়া", "দেওয়ার", "দেওয়া", "দেখতে",
    "দেখা", "দেখে", "দেন", "দেয়", "ছারা", "ধরা", "ধরে", "ধামার", "নতুন", "নয়", "না", "নাই",
    "নাকি", "নাগাদ", "নানা", "নিজে", "নিজেই", "নিজের", "নিজের", "নিত", "নিয়ে", "নিয়ে",
    "নেই", "নেওয়া", "নেওয়ার", "নেওয়া", "নয়", "পক্ষে", "পর", "পরে", "পরেই", "পরেও", "পর্যন্ত",
    "পাওয়া", "পাচ", "পারি", "পারে", "পারেন", "পি", "পোয়ে", "পেয়", "প্রতি", "প্রথম", "প্রভৃতি",
    "প্রয়ত্ত", "প্রাথমিক", "প্রায়", "প্রায়", "ফলে", "ফিরে", "ফের", "বক্তব্য", "বদলে", "বন", "বরং",
    "বলতে", "বলে", "বললেন", "বলা", "বলে", "বলেছেন", "বলেন", "বসে", "বহু", "বা", "বাদে",
    "বার", "বি", "বিনা", "বিভিন্ন", "বিশেষ", "বিশ্বয়টি", "বেশ", "বেশি", "বারবার", "ব্যাপারে", "ভাবে",
    "ভাবেই", "মতো", "মতোই", "মধ্যভাগে", "মধ্যে", "মধ্যেই", "মধ্যেও", "মনে", "মাত্র", "মাধ্যমে",
    "মোটি", "মোটেই", "যখন", "যত", "যতটা", "যাথেষ্ট", "যদি", "যদিও", "যা", "যার", "যারা",
    "যাওয়া", "যাওয়ার", "যাওয়া", "যাকে", "যাচ্ছে", "যাতে", "যাদের", "যান", "যাবে", "যাঘ", "যার",
    "যারা", "যিনি", "যে", "যেখানে", "যেতে", "যেন", "যেমন", "র", "রকম", "রয়েছে", "রাখা", "রয়ে",
    "লক্ষ", "শুধু", "শুরু", "সঙ্গে", "সঙ্গেও", "সব", "সবার", "সমস্ত", "সম্প্রতি", "সহ", "সহিত", "সাধারণ",
    "সামনে", "সি", "সুতরাং", "সে", "সেই", "সেখান", "সেখানে", "সেটা", "সেটাই", "সেটাও", "সেটি",
    "স্বয়ং", "হইতে", "হইবে", "হইয়া", "হওয়া", "হওয়ায়", "হওয়ার", "হাচ্ছে", "হত", "হতে", "হতেই", "হন",
    "হবে", "হবেন", "হয়", "হয়তো", "হয়নি", "হয়ে", "হয়েই", "হয়েছিল", "হয়েছে", "হয়েছেন", "হল",
    "হলে", "হলেই", "হলেও", "হলো", "হাজার", "হিসাবে", "হেলে", "হোক", "হয়" ]
content = [x for x in data['content'].values.tolist() if x not in stop_words]

```

Figure 9 Stop Words removal

4.4 Feature Selection and Extraction

This phase is the main part of classifying approach and this is feature selection and extraction. It actually, decides, in which perspective classify will be done. We use word count as our feature selection and create it.

4.5 Building Model and Fit Dataset for Classifier

To build a model, we separate our dataset into 3 parts - training, development and test. We will use training data to train out model and use development data to check and tune hyper parameters. And finally use test data to see how our model performs.

As, we are dealing with several classifier, we use it by importing sklearn package. This classifier can produce an integer that actually means the category of the expected news.

4.6 Accuracy of Model

For accuracy of our model we use confusion matrix, confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in dataset.

4.6.1 Decision Tree

Confusion Matrix for Decision Tree

	art-and-literature	bangladesh	durporobash	economy	\
art-and-literature	18	8	0	0	
bangladesh	5	1869	3	147	
durporobash	1	9	6	1	
economy	1	167	1	568	
education	2	23	0	10	
entertainment	3	54	3	10	
international	1	86	0	14	
life-style	7	17	0	7	
northamerica	1	8	0	2	
opinion	17	99	3	54	
sports	4	31	0	8	
technology	3	41	0	37	

	education	entertainment	international	life-style	\
art-and-literature	1	2	3	7	
bangladesh	32	50	57	11	
durporobash	1	1	3	2	
economy	10	11	27	4	
education	65	8	2	6	
entertainment	5	277	18	6	
international	4	16	151	3	
life-style	2	12	10	81	
northamerica	0	0	7	1	
opinion	17	48	39	45	
sports	8	9	12	14	
technology	2	13	19	17	

	northamerica	opinion	sports	technology
art-and-literature	0	20	9	1
bangladesh	3	84	15	39
durporobash	0	7	2	0
economy	0	42	2	46
education	0	20	2	1
entertainment	1	23	17	11
international	3	34	10	22
life-style	1	45	2	14
northamerica	5	10	0	0
opinion	8	1624	34	31
sports	0	48	486	8
technology	2	42	10	227

Figure 10 Confusion Matrix of Decision Tree

Classification Report for Decision Tree

	precision	recall	f1-score	support
art-and-literature	0.28	0.16	0.21	73
bangladesh	0.79	0.80	0.79	2276
durporobash	0.12	0.06	0.08	32
economy	0.63	0.62	0.63	909
education	0.55	0.56	0.55	128
entertainment	0.60	0.64	0.62	449
international	0.50	0.49	0.49	366
life-style	0.41	0.43	0.42	208
northamerica	0.00	0.00	0.00	34
opinion	0.81	0.82	0.81	2012
sports	0.80	0.82	0.81	600
technology	0.56	0.53	0.54	412
micro avg	0.72	0.72	0.72	7499
macro avg	0.50	0.49	0.50	7499
weighted avg	0.71	0.72	0.71	7499

Figure 11 Classification Report for Decision Tree

Accuracy for Decision Tree 0.72

4.6.3 Random Forest

Confusion Matrix		for Random Forest			
	art-and-literature	bangladesh	durporobash	economy	\
art-and-literature	3	7	0	0	
bangladesh	0	2148	0	34	
durporobash	0	18	2	1	
economy	0	209	0	586	
education	0	41	0	13	
entertainment	0	84	0	3	
international	0	200	0	7	
life-style	0	21	0	4	
northamerica	0	15	0	0	
opinion	0	71	0	3	
sports	0	14	0	1	
technology	0	70	0	34	
	education	entertainment	international	life-style	\
art-and-literature	0	4	0	0	
bangladesh	0	4	0	0	
durporobash	0	1	0	0	
economy	0	0	1	0	
education	47	4	0	0	
entertainment	0	295	0	0	
international	0	5	60	0	
life-style	0	4	0	65	
northamerica	0	0	0	0	
opinion	0	0	0	0	
sports	0	7	0	0	
technology	0	2	0	0	
	northamerica	opinion	sports	technology	
art-and-literature	0	54	1	0	
bangladesh	0	125	3	1	
durporobash	0	11	0	0	
economy	0	75	0	8	
education	0	33	1	0	
entertainment	0	42	4	0	
international	0	66	4	2	
life-style	0	103	1	0	
northamerica	3	15	1	0	
opinion	0	1942	3	0	
sports	0	23	582	1	
technology	0	84	2	221	

Figure 12 Confusion Matrix for Random Forest

Classification Report for Random Forest

	precision	recall	f1-score	support
art-and-literature	0.00	0.00	0.00	73
bangladesh	0.74	0.93	0.82	2276
durporobash	1.00	0.06	0.12	32
economy	0.85	0.66	0.74	909
education	0.98	0.48	0.64	128
entertainment	0.88	0.65	0.75	449
international	1.00	0.17	0.29	366
life-style	0.99	0.37	0.54	208
northamerica	0.00	0.00	0.00	34
opinion	0.76	0.97	0.85	2012
sports	0.96	0.93	0.95	600
technology	0.95	0.54	0.69	412
micro avg	0.79	0.79	0.79	7499
macro avg	0.76	0.48	0.53	7499
weighted avg	0.81	0.79	0.77	7499

Figure 13 Classification Report for Random Forest

Accuracy for Random Forest 0.81

4.6.4 Naive Bayes Classifier

Confusion Matrix for Naive Bayes

	art-and-literature	bangladesh	durporobash	economy	\
art-and-literature	64	0	0	0	
bangladesh	12	1968	2	103	
durporobash	9	4	13	1	
economy	1	40	0	783	
education	3	16	0	4	
entertainment	4	8	0	0	
international	0	23	1	12	
life-style	15	1	0	0	
northamerica	2	1	0	1	
opinion	76	126	0	82	
sports	0	3	0	1	
technology	1	5	0	17	

	education	entertainment	international	life-style	\
art-and-literature	0	1	0	0	
bangladesh	22	33	26	14	
durporobash	0	2	2	1	
economy	7	5	10	2	
education	96	8	0	2	
entertainment	0	406	5	1	
international	0	13	279	0	
life-style	4	8	0	160	
northamerica	0	1	16	0	
opinion	11	8	69	25	
sports	0	10	7	0	
technology	1	0	10	2	

	northamerica	opinion	sports	technology
art-and-literature	0	3	1	0
bangladesh	0	113	4	18
durporobash	0	1	0	0
economy	0	5	1	25
education	0	6	1	3
entertainment	0	2	2	0
international	0	7	3	6
life-style	0	7	0	3
northamerica	2	8	1	2
opinion	0	1600	8	14
sports	0	1	605	1
technology	0	3	1	373

Figure 14 Confusion Matrix for Naive Bayes

Classification	Report for Naive Bayes			
	precision	recall	f1-score	support
art-and-literature	0.34	0.86	0.48	73
bangladesh	0.90	0.85	0.87	2276
durporobash	0.92	0.34	0.50	32
economy	0.77	0.89	0.82	909
education	0.64	0.77	0.70	128
entertainment	0.84	0.93	0.88	449
international	0.75	0.85	0.80	366
life-style	0.74	0.82	0.78	208
northamerica	1.00	0.09	0.16	34
opinion	0.92	0.81	0.86	2012
sports	0.97	0.97	0.97	600
technology	0.84	0.90	0.87	412
micro avg	0.85	0.85	0.85	7499
macro avg	0.80	0.76	0.73	7499
weighted avg	0.87	0.85	0.85	7499

Figure 15 Classification report for Naive Bayes

Accuracy for Naive Bayes 0.8518

4.7 Compare Precision

Algorithm	Accuracy
Decision Tree	0.71
Random Forest	0.81
Naive Bayes	0.87

Table 2 Compare Precision

4.8 Compare Recall

Algorithm	Accuracy
Decision Tree	0.72
Random Forest	0.79
Naive Bayes	0.85

Table 3 Compare Recall

4.9 F1 Score for all algorithms

Algorithm	Accuracy
Decision Tree	0.71
Random Forest	0.77
Naive Bayes	0.85

Table 4 F1 Score for all algorithm

4.10 Compare Algorithm Accuracy

Algorithm	Accuracy
Decision Tree	71%
Random Forest	81%
Naive Bayes	85%

Table 5 Compare Algorithm Accuracy

From the above comparison tables, we see that, in the case of Precision, Recall, f1-score and accuracy Naive Bayes classifier is the best. The values of precision, recall and f1-score are respectively 0.87, 0.85, and 0.85 and the accuracy of this model is 85% that is highest value in comparison with all classifier.

4.10 Summary

After getting this accuracy, highest result come from Naive Byes that's why, we are satisfied, if we are try to increase accuracy level, must to prepare the dataset properly. The all categorical news should be equally numbered. At that, to increase the accuracy level, data cleaning has not alternative. The more data are preprocessed, the more accurate prediction will be shown by this classifier.

Chapter 5

Conclusion and Future Direction

5.1 Introduction

Summary of the Study It has no doubt that there are lots of research works on Natural Language Processing especially on English Language. When the outcome of such kind of work is taking a revolutionary change in our computing life, recently, such kind of research is being increased this time. We get some outstanding real life applications on the blessing of such kind of research works. But it is a matter of great regret that there has no such of research work on Bangla Language. But it is the hope for us that many of researchers from various countries have started to do research on this field. In our research work, we do some approaches of our Bangla News to classify its category.

5.2 Conclusion

Though, the accuracy level of the classifier algorithm that we used in our project is not so good but we have learnt lots of things from this research. We can now deal with the Bangla Text. We can now preprocess the row data. And can apply the classifier on our trained dataset. Hope, it will be very beneficial to the future researchers to do such kind of research on Bangla Text or Bangla news.

5.3 Recommendations

A few notable recommendations for this are as follows

- ∄ To create the data set more efficiently, can produce a better output of this research work.
- ∄ Applying some deep leaning method could be improve the model performance

5.4 Future Direction

- ∄ Adding more categories in this project, can make this more efficient.
- ∄ Using more classifiers on this dataset, can get a better understanding on which classifier can be the best for this work.

References

1. Accuracy, Precision, Recall or F1 <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>
2. Machine Learning Classifiers <https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623>
3. Types of classification algorithms in Machine Learning <https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14>
4. The 10 Best Machine Learning Algorithms for Data Science Beginners <https://www.dataquest.io/blog/top-10-machine-learning-algorithms-for-beginners/>
5. Getting started with Data Analysis with Python Pandas <https://towardsdatascience.com/getting-started-to-data-analysis-with-python-pandas-with-titanic-dataset-a195ab043c77>
6. Exploratory Data Analysis with Pandas <https://www.kaggle.com/kashnitsky/topic-1-exploratory-data-analysis-with-pandas>
7. Data analysis using Pandas <https://www.geeksforgeeks.org/python-data-analysis-using-pandas/>
8. How to Scrape the Web using Python with Scrapy Spiders <https://towardsdatascience.com/how-to-scrape-the-web-using-python-with-scrapy-spiders-e2328ac4526>
9. Writing a Simple Web Scraper using Scrapy <https://www.codementor.io/andy995/writing-a-simple-web-scraper-using-scrapy-myb7vrmgx>
10. Web Scraping in Python using Scrapy (with multiple examples) <https://www.analyticsvidhya.com/blog/2017/07/web-scraping-in-python-using-scrapy/>
11. Support - http://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html
12. F1-score - http://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html