DEERWALK INSTITUTE OF TECHNOLOGY

Tribhuvan University

Institute of Science and Technology

NEPALI NEWS CLASSIFICATION USING NAÏVE BAYES

A PROJECT REPORT

Submitted to

Department of Computer Science and Information Technology

DWIT College

In partial fulfillment of the requirements for the Bachelor's Degree in Computer Science and Information Technology

Submitted by

Arika Joshi, Iris Raj Pokhrel

June, 2019

DWIT College

DEERWALK INSTITUTE OF TECHNOLOGY

Tribhuvan University

SUPERVISOR'SRECOMENDATION

I hereby recommend that this project prepared under my supervision by ARIKA JOSHI and IRIS RAJ POKHREL entitled **"NEPALI NEWS CLASSIFICATION USING NAÏVE BAYES"** in partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Information Technology be processed for the evaluation.

…………………………………………

Dr. Sunil Chaudhary

HOD, Department of Computer Science

Deerwalk Institute of Technology

DWIT College

DWIT College

DEERWALK INSTITUTE OF TECHNOLOGY

Tribhuvan University

LETTER OF APPROVAL

This is to certify that this project prepared by ARIKA JOSHI and IRIS RAJ POKHREL entitled **“NEPALI NEWS CLASSIFICATION USING NAÏVE BAYES”** in partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Information Technology has been well studied. In our opinion it is satisfactory in the scope and quality as a project for the required degree.

…………………………………… Dr. Sunil Chaudhary [Supervisor] HoD, Department of Computer Science DWIT College	………………………………………… Mr. Hitesh Karki Chief Academic Officer DWIT College
………………………………………….. Dr. Subarna Shakya Professor, Department of Electronics and Computer Engineering Pulchowk Campus, IOE, Tribhuvan University	………………………………………….. Mr. Ritu Raj Lamsal HoD, Department of Electronics DWIT College

ACKNOWLEDGEMENT

We would like to thank Dr. Sunil Chaudhary, HOD, Department of Computer Science, DWIT College for supervising this project and giving assistance when faced with problems. This project would not have been possible without his guidance and persistent help. We would like to thank him for providing deeper knowledge about the technology needed for this project.

We would also like to thank Mr. Hitesh Karki, Chief Academic Officer, DWIT College and Mr. Ritu Raj Lamsal, HOD, Department of Electronics, DWIT College for providing help with various aspect of this project.

Last but not the least, our sincere thanks go to our parents and family members, who have always supported us and to all our friends who directly or indirectly helped us to complete this project.

Arika Joshi (TU Roll Number: 7439/072)

Iris Raj Pokhrel (TU Roll Number: 7448/072)

Date: - June, 2019

ABSTRACT

Online news portal and other media on the internet now produce the large amount of text, which is mostly unstructured in nature. When an individual wants to access or share particular news, it should be organized or classified in the proper class. Nepali news classification is the task of categorizing the news content into the predefined category from the training news dataset.

In this project, a system has been built for categorizing the content of the news into different categories using the news article from a major Nepali language newspaper published in Nepal. This project evaluates some widely used machine learning techniques mainly Naïve Bayes for automatic Nepali News classification problem. It classifies the news by analyzing the content of the news.

In the system, a self-created Nepali News Corpus with 16 different categories and total 16719 documents collected by crawling different online national news portal is used. The test showed the accuracy of 83.94% in the news classification using Naïve Bayes.

In the system, the user can categorize the news based on their content by analyzing the news content from various Nepali language newspaper and clicking the categorize button.

Keywords: *News Classification; Naïve Bayes Classifier; Text Mining; Text Processing*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVATIONS

CPM	Critical Path Model
CSS	Cascading Style Sheet
HTML	Hypertext Markup Language
IDF	Inverse Document Frequency
PHP	Hypertext Preprocessor
TF	Term Frequency
TFID	Term Frequency-Inverse Document Frequency
UI	User Interface
SVM	Support Vector Machine
SAMNews	Sentiment mining for Malay Newspaper

LIST OF SYMBOLS

Symbol	Name	Definition
%	Percentage	Percentage of the sample
*	Multiplication	Product of two terms
/	Division	Division of two products

CHAPTER 1: INTRODUCTION

1.1 Overview

Nowadays on the Internet there are a lot of sources that generate immense amounts of daily news. In addition, the demand for information by users has been growing continuously, so it is crucial that the news is classified to allow users to access the information of interest quickly and effectively. There exists a large amount of information being stored in the electronic format. With such data, it has become a necessity of such means that could interpret and analyze such data and extract such facts that could help in decision making

Classification is quite a challenging field as it requires prepossessing steps to convert unstructured data to structured information. Nepali News Classification automatically predicts the incoming news type to some of the predefined classes using the trained classifier. It classifies the news on the basis of their content. Nepali News Classification offers news to the public with the added categorization on the basis of the content of the news. Thus, the aim is to build models that take input as news content and output as news category.

1.2 Background and Motivation

News classification is a growing interest in the research of text mining. Correctly identifying the news into particular category is still presenting challenge because of large and vast number of features in the dataset. In regards to the existing classifying approaches, Naïve Bayes' is potentially good at serving as a document classification model because Naïve Bayes model is very simple and is also potentially good due to its simplicity.

News Classification is the task in which sorting is done automatically to classify the documents into predefined classes. Manual news classification is an expensive and time-consuming method, as it become difficult to classify millions of documents manually. Therefore, Nepali News classification is constructed using labeled documents and its accuracy is much better than manual text classification and it is less time consuming too. The system includes the use of Naïve Bayes for news classification. In the proposed work 16 different types of news has been classified like business, sports, entertainment, political, literature and many more.

1.3 Problem Statement

News information was not easily and quickly available until the beginning of last decade. But now news is easily accessible via content providers such as online news services. A huge amount of information exists in form of text in various diverse areas whose analysis can be beneficial in several areas. The task of manually labeling the news class becomes tedious when a large amount of news comes together from different sources. It is almost impossible to make the classification manually if some application tries to feed the trending news to the reader in real time. Hence it is necessary to develop an automatic tool that will be able to classify the Nepali news into relevant class.

Classification is quite challenging field as it requires preprocessing steps to convert unstructured data to structured information. With the increase in the number of news it has got difficult for users to access news of his interest which makes it necessity to categories news so that it could be easily accessed. When it comes to news it is much difficult to classify as news are continuously appearing that need to be processed and that news could never be seen before and could fall in a new category.

1.4 Objective of the Project

1.4.1 General Objective

- Develop an automatic tool that will be able to classify the Nepali news into relevant class.
- Automatically predict the incoming news type to some of the predefined classes using the trained classifier.

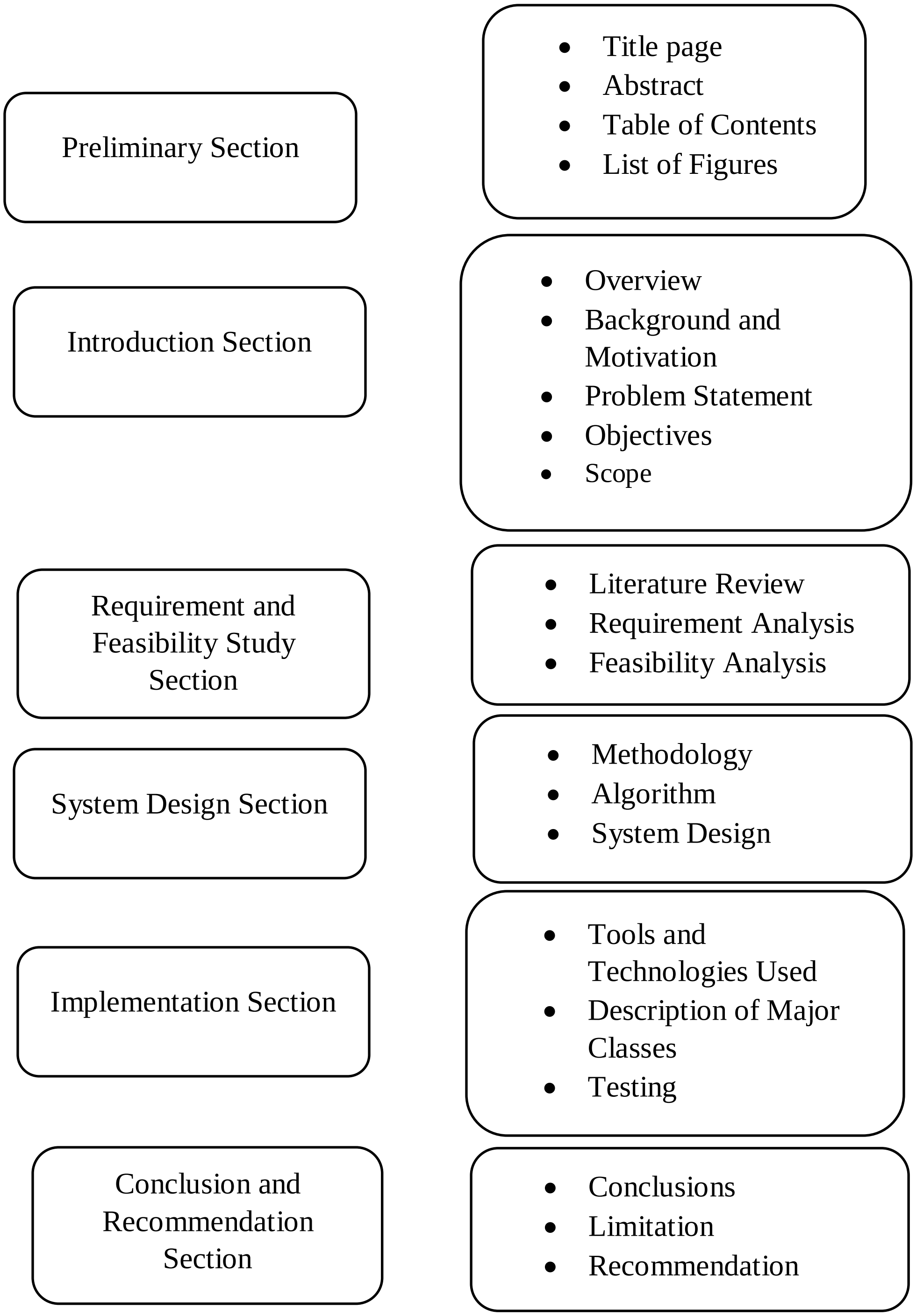
1.4.2 Specific Objective

- To implement Natural Language Processing, Regular Expression and concepts of Document Object Model.
- To crawl different online national news portals for content-based news classification.
- To classify the news on the basis of the content.

1.5 Scope of the Project

News Classification offers news to the public with the added categorization on the basis of the content of the news. It can be used by the general public to know under which category the news fall.

1.6 Outline of the Report



Preliminary Section	• Title page • Abstract • Table of Contents • List of Figures
Introduction Section	• Overview • Background and Motivation • Problem Statement • Objectives • Scope
Requirement and Feasibility Study Section	• Literature Review • Requirement Analysis • Feasibility Analysis
System Design Section	• Methodology • Algorithm • System Design
Implementation Section	• Tools and Technologies Used • Description of Major Classes • Testing
Conclusion and Recommendation Section	• Conclusions • Limitation • Recommendation

Figure 1:Outline of the Document

CHAPTER 2: REQUIREMENT AND FEASIBILITY ANALYSIS

2.1 Literature Review

2.1.1 Sentiment Mining of Malay Newspaper (SAMNews) Using Artificial Immune System

There is sheer volume of rich web resources such as digital newspaper, e-forum, blogs, Facebook and Twitter. Mining the digital text resources may reveal interesting knowledge to respective individuals or organizations. Text mining and sentiment mining or analysis are parts of a new area in sentiment research.SAMNews is constructed based on the artificial immune system called negative selection algorithm which is able to classify the sentiment in newspaper's sentences into the polarity (positive, negative or neutral) intelligently. [1]. The sentiment analysis in this project utilized 1000 sentences from newspapers to evaluate the average accuracy. The research used 900 sentences from newspapers as the training data and another 100 as the testing data. The accuracy is achieved at 88.5%. In the future, a comparative study on Artificial Immune System and other techniques or algorithms can be carried out to enhance the performance of the sentiment classification model.

Our project being classification of news based on content and classifying on the basis of categories and altogether of 16 categories were initially used in our project. The above SAM News were limited to only three categories (Politics, natural disaster and economy).

Hence, some references were taken from the above research paper about the similar work like ours but their idea and prospect being limited than ours, we couldn't adopt every methodologies and concept used here.

2.1.2 News Classification using Support Vector Machine

Digital news with a variety topic is abundant on the internet. The problem is to classify news based on its appropriate category to facilitate user to find relevant news rapidly. Classifier engine is used to split any news automatically into the respective category. This research employs SVM to classify Indonesian news. SVM is a robust method to classify binary classes. The core processing of SVM is in the formation of an optimum separating plane to separate the different classes. For multiclass problem, a mechanism called one against one is used to combine the binary classification result.

The Indonesian news classification using SVM [2] helped us in designing the classification system for our project. They used three methodologies in the project; Input Data (News Document), Preprocessing (Tokenizing, stop word Removal, Stemming, Text Frequency) and Output Data (Learning and Classification). We used the similar methodology in our Nepali News Classification using Naïve Bayes. The journal [2], helped us in creating a proper framework to the project and hence became a backbone/reference throughout our project.

2.1.3 News Classification and Its Techniques

Text mining has gained quite a significant importance during the past few years. Data, now-a-days is available to users through many sources like electronic media, digital media and many more. This data is usually available in the most unstructured form and there exists a lot of ways in which this data may be converted to structured form. In many real-life scenarios, it is highly desirable to classify the information in an appropriate set of categories. News contents are one of the most important factors that have influence on various sections.

2.2 Requirement Analysis

2.2.1 Functional Requirement

The functional requirements of the project are listed as follows:

- The user shall be able to classify the news article into predefined categories.
- The user shall be going through the news, analyze the content, classify it into the proper class and view the results.

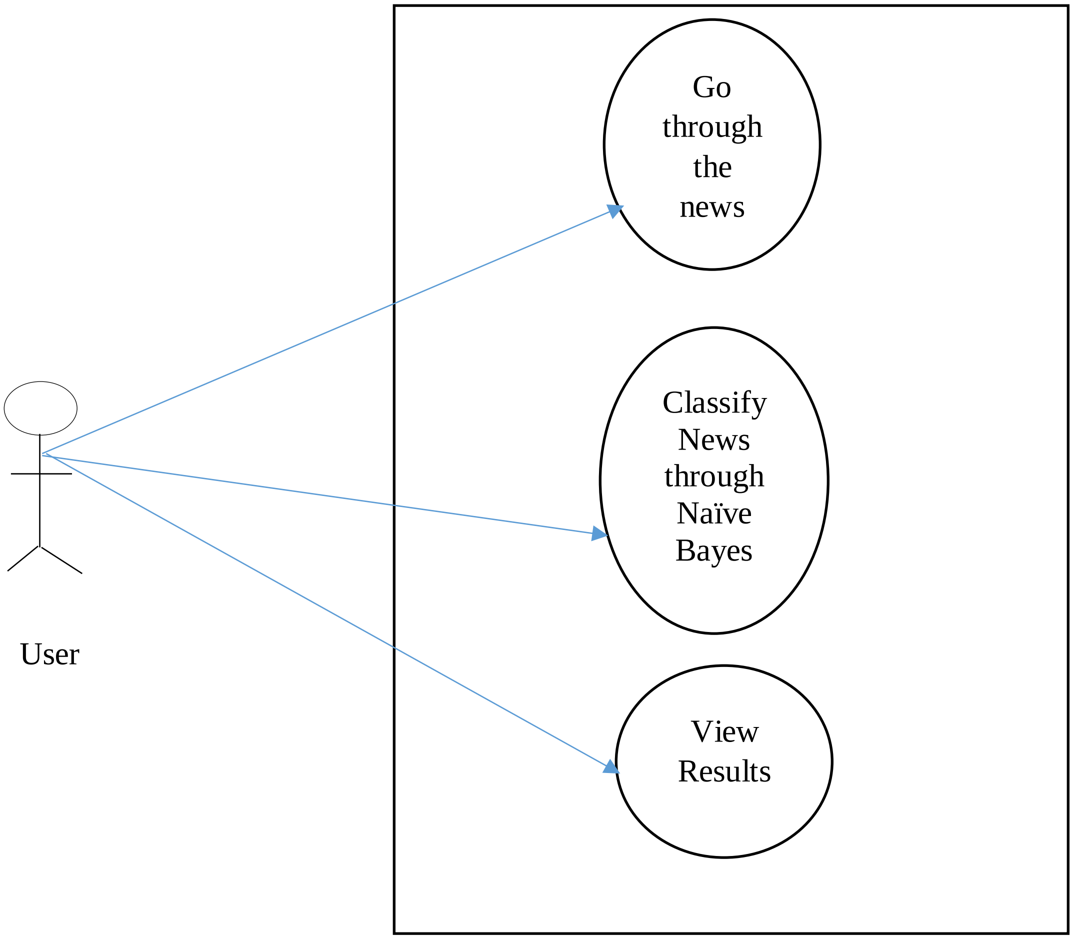


Figure 2: Use Case Diagram

2.2.2 Non-Functional Requirement

The non-functional requirements of the project are listed as follows:

- All news must be legitimate news.
- Websites from which data is to be extracted must be crawler friendly.
- The Nepali News classification must have a clear and easily maintainable interface for classifying the news.
- Operating of the application must not be time and resource consuming.
- It must be cheap to maintain.

2.3 Feasibility Analysis

After gathering of the required resources, whether the completion of the project with the gathered resource is feasible or not is checked using the following feasibility analysis. In order to test the feasibility of the system, four different studies have been carried out and they are:

2.3.1 Technical Feasibility

The project is technically feasible as it can be built using the existing available technologies. The tools, modules and libraries needed to build the system are open source, freely available and are easy to use. Furthermore, there are enough resources which are being produced every year. These human skills can be implemented for the conduction, management and maintenance. Nepali News Classification uses Flask framework which is a python-based framework. JavaScript, CSS and HTML pages has been used for development of front end while Python is used in backend.

2.3.2 Economic Feasibility

Developing an IT application is an investment. Since after developing that application it provides the organization with profits. Profits can be monetary or in the form of an improved working environment. However, it carries risks, because in some cases an estimate can be wrong. And the project might not actually turn out to be beneficial. Cost benefit analysis helps to give management a picture of the costs, benefits and risks. It usually involves comparing alternate investments. Implementing this system is a huge save in financial point as time is equivalent to money.

2.3.3 Operational Feasibility

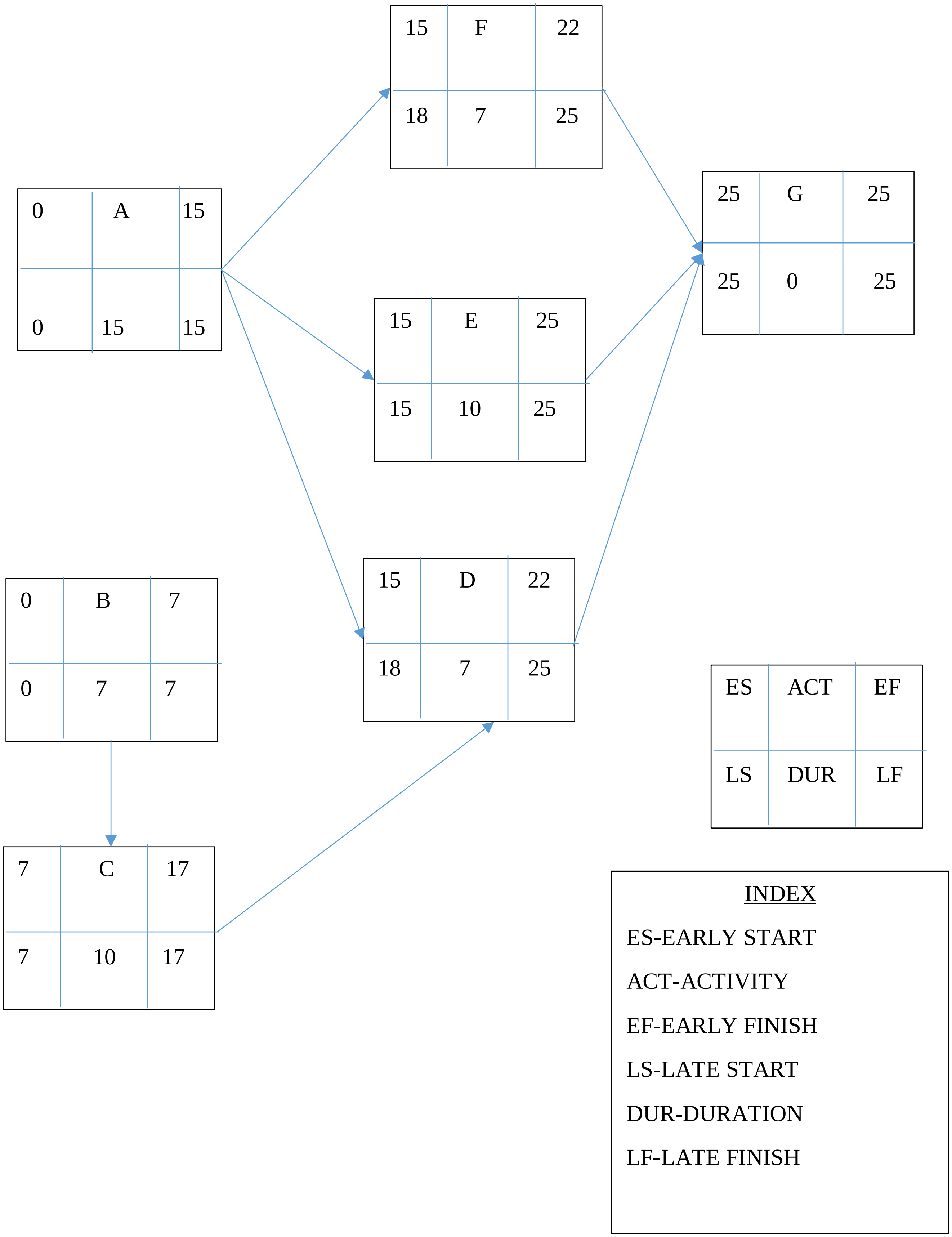
Nepali News Classification has a simple design and is easy to use. The web-based application can operate in a system having Windows, Linux or MacOS. It uses two-tier architecture (i.e. Client and Server). It can be easily accessed and can be used to classify news on the basis of their content into different categories. The resource-intensive part of the project was training the data set. Considering the above cases, the project is operationally feasible.

2.3.4 Schedule Feasibility

The schedule feasibility analysis is carried out using the CPM method. With CPM, critical tasks were identified and interrelationship between tasks were identified which helped in planning that defines critical and non-critical tasks with the goal of preventing time-frame problems and process bottlenecks.

Table 1:Critical Path Method Analysis

ACTIVITY	TIME(DAYS)	PREDECESSOR
Research on previous work and papers(A)	15	
Classification Model Design(B)	7	-
News Collection(C)	10	B
Naïve Bayes' Algorithm Implementation(D)	7	A, C
UI Design and Testing Phase(E)	10	A, D
Documentation(F)	7	A



INDEX

ES-EARLY START

ACT-ACTIVITY

EF-EARLY FINISH

LS-LATE START

DUR-DURATION

LF-LATE FINISH

Figure 3:Critical Path Method Analysis

CHAPTER 3: METHODOLOGY

The waterfall development model was followed for this project. For instance, a crawler was made and it helped in data collection. The collected data was the input for making a data set for the application. Since all the model phases were processed and completed one at a time, overlapping challenges were not encountered. process. It was used because the model phases are processed and completed one at a time and these phases do not overlap.

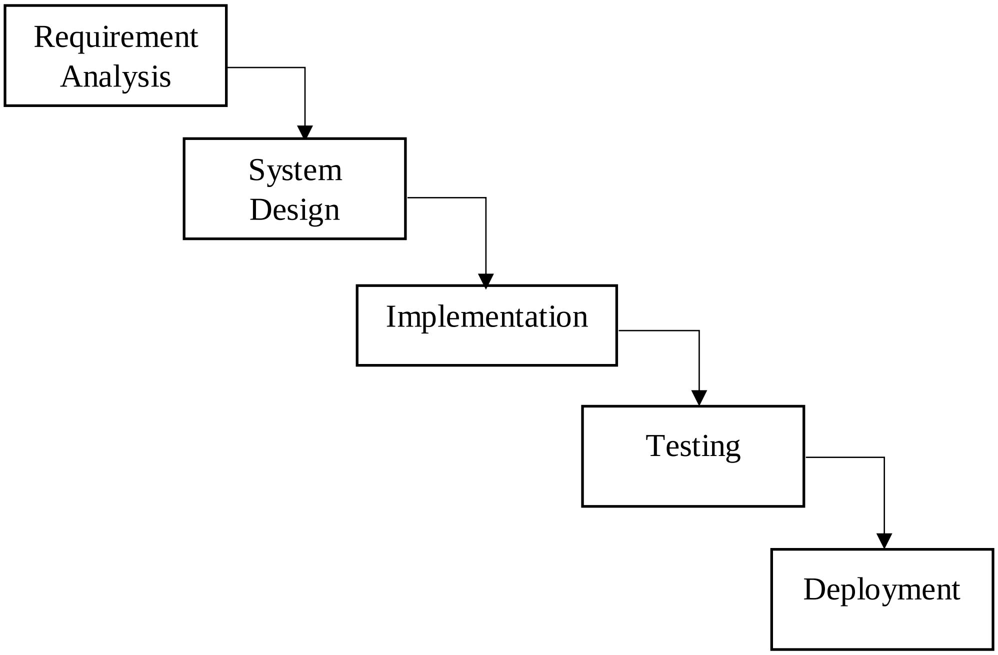


Figure 4:Implementation of Waterfall Model.

Figure 4 represents the waterfall development methodology which was used in the development of the project. Firstly, requirement analysis is carried out. In this news is accumulated from various sources. A collection of Nepali News was collected from various online Nepali News portals.

For system design and implementation, it uses Flask framework which is a python-based framework. JavaScript, CSS and HTML pages has been used for development of front end while

Python is used in backend. It includes preprocessing input data on the form of categorized news including tokenization, stop word filtering and stemming.

For testing and deployment, three cases: cross validating the trained data, accuracy of the system and overall system testing are included. With these it develops an automatic tool that will be able to classify the Nepali news into relevant class.

3.1 Data Preparation

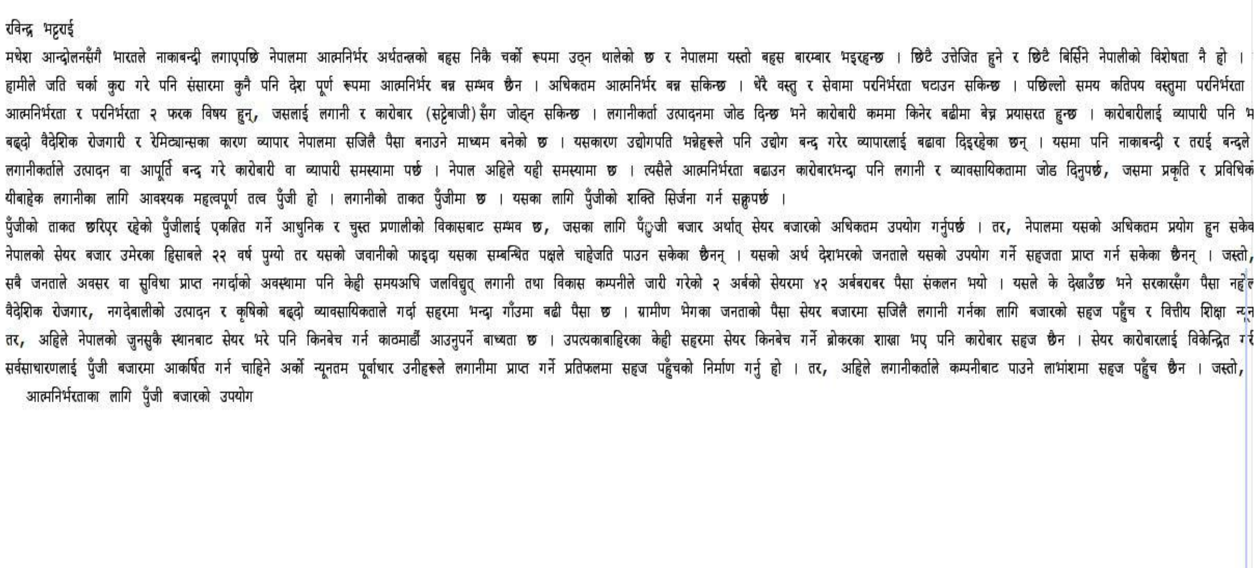
3.1.1 Data Collection

The Nepali language belongs to one of the most common scripts, Devanagari, invented by Brahmins around the 11th century. It consists of 36 consonant symbols, 12 vowel symbols and 10 numeral symbols.

The first step of news classification is accumulating news from various sources. This data may be available from various sources like newspapers, press, magazines, world wide web and many more. But with the widespread network and information technology growth internet has emerged as the major source for obtaining news. Data may be available in a various format.

Table 2:News Class and Total Number of Documents

S. N	News Class	Number of Documents
1	Auto	95
2	Bank	351
3	Blog	209
4	Business Interview	142
5	Economy	1188
6	Education	85
7	Entertainment	1174
8	Interview	87
9	Literature	16
10	National News	7452
11	Opinion	590
12	Sports	2285
13	Technology	110
14	Tourism	214
15	World	212
16	Employment	154



रविन्द्र भट्टराई
मधेश आन्दोलनसँगै भारतले नाकाबन्दी लगाएपछि नेपालमा आत्मनिर्भर अर्थतन्त्रको बहस निकै चर्को रूपमा उठ्न थालेको छ र नेपालमा यस्तो बहस बारम्बार भइरहन्छ । छिटै उत्तेजित हुने र छिटै बिर्सिने नेपालीको विशेषता नै हो ।
हामीले जति चर्का कुरा गरे पनि संसारमा कुनै पनि देश पूर्ण रूपमा आत्मनिर्भर बन्न सम्भव छैन । अधिकतम आत्मनिर्भर बन्न सकिन्छ । धेरै वस्तु र सेवामा परनिर्भरता घटाउन सकिन्छ । पछिल्लो समय कतिपय वस्तुमा परनिर्भरता
आत्मनिर्भरता र परनिर्भरता २ फरक विषय हुन्, जसलाई लगानी र कारोबार (सट्टेबाजी) सँग जोड्न सकिन्छ । लगानीकर्ता उत्पादनमा जोड दिन्छ भने कारोबारी कममा किनेर बढीमा बेच्न प्रयासरत हुन्छ । कारोबारीलाई व्यापारी पनि भ
बढ्दो वैदेशिक रोजगारी र रेमिट्यान्सका कारण व्यापार नेपालमा सजिलै पैसा बनाउने माध्यम बनेको छ । यसकारण उद्योगपति भन्नेहरूले पनि उद्योग बन्द गरेर व्यापारलाई बढावा दिइरहेका छन् । यसमा पनि नाकाबन्दी र तराई बन्दले
लगानीकर्ताले उत्पादन वा आपूर्ति बन्द गरे कारोबारी वा व्यापारी समस्यामा पर्छ । नेपाल अहिले यही समस्यामा छ । त्यसैले आत्मनिर्भरता बढाउन कारोबारभन्दा पनि लगानी र व्यावसायिकतामा जोड दिनुपर्छ, जसमा प्रकृति र प्रविधिक
यीबाहेक लगानीका लागि आवश्यक महत्वपूर्ण तत्व पुँजी हो । लगानीको ताकत पुँजीमा छ । यसका लागि पुँजीको शक्ति सिर्जना गर्न सक्नुपर्छ ।
पुँजीको ताकत छरिएर रहेको पुँजीलाई एकत्रित गर्ने आधुनिक र चुस्त प्रणालीको विकासबाट सम्भव छ, जसका लागि पँ◌ुजी बजार अर्थात् सेयर बजारको अधिकतम उपयोग गर्नुपर्छ । तर, नेपालमा यसको अधिकतम प्रयोग हुन सकेव
नेपालको सेयर बजार उमेरका हिसाबले २२ वर्ष पुग्यो तर यसको जवानीको फाइदा यसका सम्बन्धित पक्षले चाहेजति पाउन सकेका छैनन् । यसको अर्थ देशभरको जनताले यसको उपयोग गर्ने सहजता प्राप्त गर्न सकेका छैनन् । जस्तो,
सबै जनताले अवसर वा सुविधा प्राप्त नगर्दाको अवस्थामा पनि केही समयअघि जलविद्युत् लगानी तथा विकास कम्पनीले जारी गरेको २ अर्बको सेयरमा ४२ अर्बबराबर पैसा संकलन भयो । यसले के देखाउँछ भने सरकारसँग पैसा नहो
वैदेशिक रोजगार, नगदेबालीको उत्पादन र कृषिको बढ्दो व्यावसायिकताले गर्दा सहरमा भन्दा गाँउमा बढी पैसा छ । ग्रामीण भेगका जनताको पैसा सेयर बजारमा सजिलै लगानी गर्नका लागि बजारको सहज पहुँच र वित्तीय शिक्षा न्यून
तर, अहिले नेपालको जुनसुकै स्थानबाट सेयर भरे पनि किनबेच गर्न काठमाडौं आउनुपर्ने बाध्यता छ । उपत्यकाबाहिरका केही सहरमा सेयर किनबेच गर्ने ब्रोकरका शाखा भए पनि कारोबार सहज छैन । सेयर कारोबारलाई विकेन्द्रित गर
सर्वसाधारणलाई पुँजी बजारमा आकर्षित गर्न चाहिने अर्को न्यूनतम पूर्वाधार उनीहरूले लगानीमा प्राप्त गर्ने प्रतिफलमा सहज पहुँचको निर्माण गर्नु हो । तर, अहिले लगानीकर्ताले कम्पनीबाट पाउने लाभांशमा सहज पहुँच छैन । जस्तो,
आत्मनिर्भरताका लागि पुँजी बजारको उपयोग

Figure 5: Data Set Preprocessing on IDE

3.1.2 Data Selection/Filtering

After the collection of news text pre-processing is done. As this data comes from variety of data gathering sources and its cleaning is required so that it could be free from all corrupt and futile data. Data now needs to be discriminated from unrelated words like semicolon, commas, double quotes, full stop, and brackets, special characters etc. Data is made free from those words which appear customarily in text and are known as stop words.

News tokenization involves fragmenting the huge text into small tokens. Each word in the news is treated as a string. The output of this step is treated as input for the next steps involved in text mining.

The stop words language is specific and does not carry any information. It generally includes conjunctions, pronoun and prepositions. They are contemplated of low worth and are removed eventually. These words need to be percolate before the processing of data. Stop words can be removed from data in many ways. These removals can be on the basis of concepts i.e. the removal will be of the words which provide very fewer information about classification. Stop words are high-frequency words that has not much influence in the text are removed to increase the performance of the classification. The list of 255 stop-words like “छ, म, हो, केह, हामी, मेरो, यो, ह, फेर, आफू, हुछ, राख, भयो, गनु, पन, etc.” were collected and removed from the text.

After the removal of stop words the next activity that is performed is stemming. This step reduces a word to its root. The motive behind using stemming is to remove the suffixes so that the number of words would be brought down. Stemming is used to reduce the given word into its stem. Since the word stem reflects the meaning of a particular word, we have segmented the inflected word and derivational word into a stem word so that the dimension of vocabulary reduced in the significant manner.

The text preprocessing cleans the text data to make it ready to use in training and testing of the machine learning model. Preprocessing is done to reduce the noise in the text that helps to improve the performance of the classifier and speed up the classification process, thus aiding in real time news classification. The main preprocessing techniques used are again explained below.

1. Tokenization: Breakdowns the text into sentences and then words. Vertical bar, question mark, and full stop are used to break down the sentences and while space and comma are used to break down the words.
2. Special symbol and number removal: Special symbols and numbers, those do not have much importance in classification, are removed.
3. Stop word removal: Stop words are high-frequency words that has not much influence in the text are removed to increase the performance of the classification.
4. Word Stemming: Stemming is used to reduce the given word into its stem.

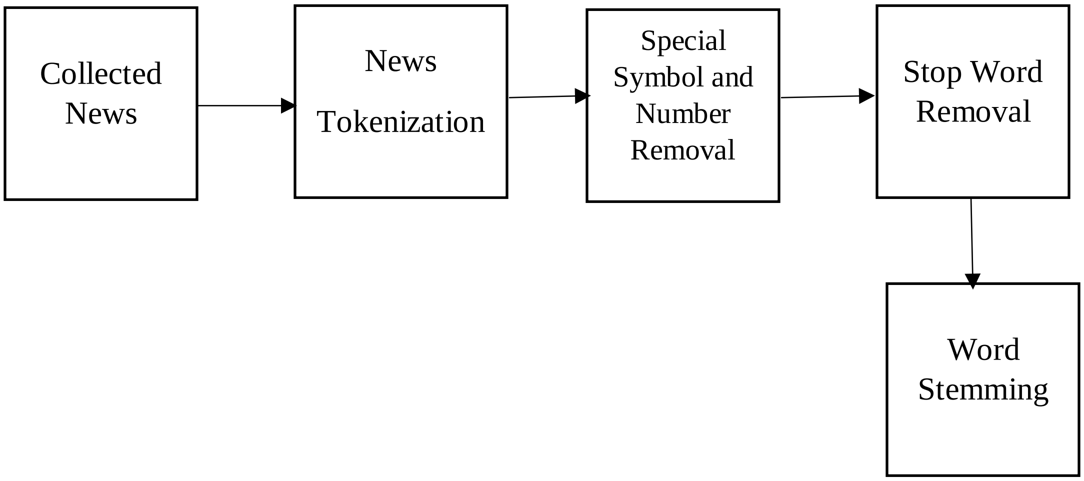


Figure 6:News Preprocessing System

Feature Vector construction is the process of representing the news into a vector form. To represent Nepali news in vector form, the TF-IDF weighting value for each word in the text is taken as a dimensional value in a vector. It is calculated as,

$W_{t, d, D} = tf_{t, d} * idf_{t, D}$ ………………………………. eq (A) where,

$tft, d = ft, d / \max \{ft', d: t' \in d\}$ …………………… eq (1)

$idft, D = \log 1+ |\{d \in {}^{N}D: t \in d\}_{|}$……………………. eq (2)

Here,

eq(A)= eq (1) * eq (2) ………………………eq (B)

tf = Term frequency. idf = Inverse document frequency. $f_{t, d}$ = Number of term t in document d. max $\{f_{t', d} : t' \in d\}$ = Max occurring term t’ in document d.

N = Total number of documents in the corpus D. $|\{d \in D: t \in d\}|$= Number of documents where the term t appears.

```
tfidfVectorizer = TfidfVectorizer(tokenizer=lambda x: x.split(" "),
                                  sublinear_tf=True,encoding='utf-8',
                                  decode_error='ignore',
                                  stop_words=stopWords)

vectorized = tfidfVectorizer.fit_transform(xTrain)
with open('tfidf.pkl', 'wb') as f:
    dill.dump(tfidfVectorizer, f)
print('No of Samples , No. of Features ', vectorised.shape)
clf1 = Pipeline([
    ('vect', tfidfVectorizer),
('clf', MultinomialNB(alpha=0.01, fit_prior=True))
])
```

Figure 7:Implementation of Tfidf Vectorizer

Naive Bayes Classifier is a simple probabilistic classifier based on Bayes Theorem with strong independence assumptions of feature space. Depending on the precise nature of the probability model, Naive Bayes classifier can be trained very efficiently in a supervised learning setting.

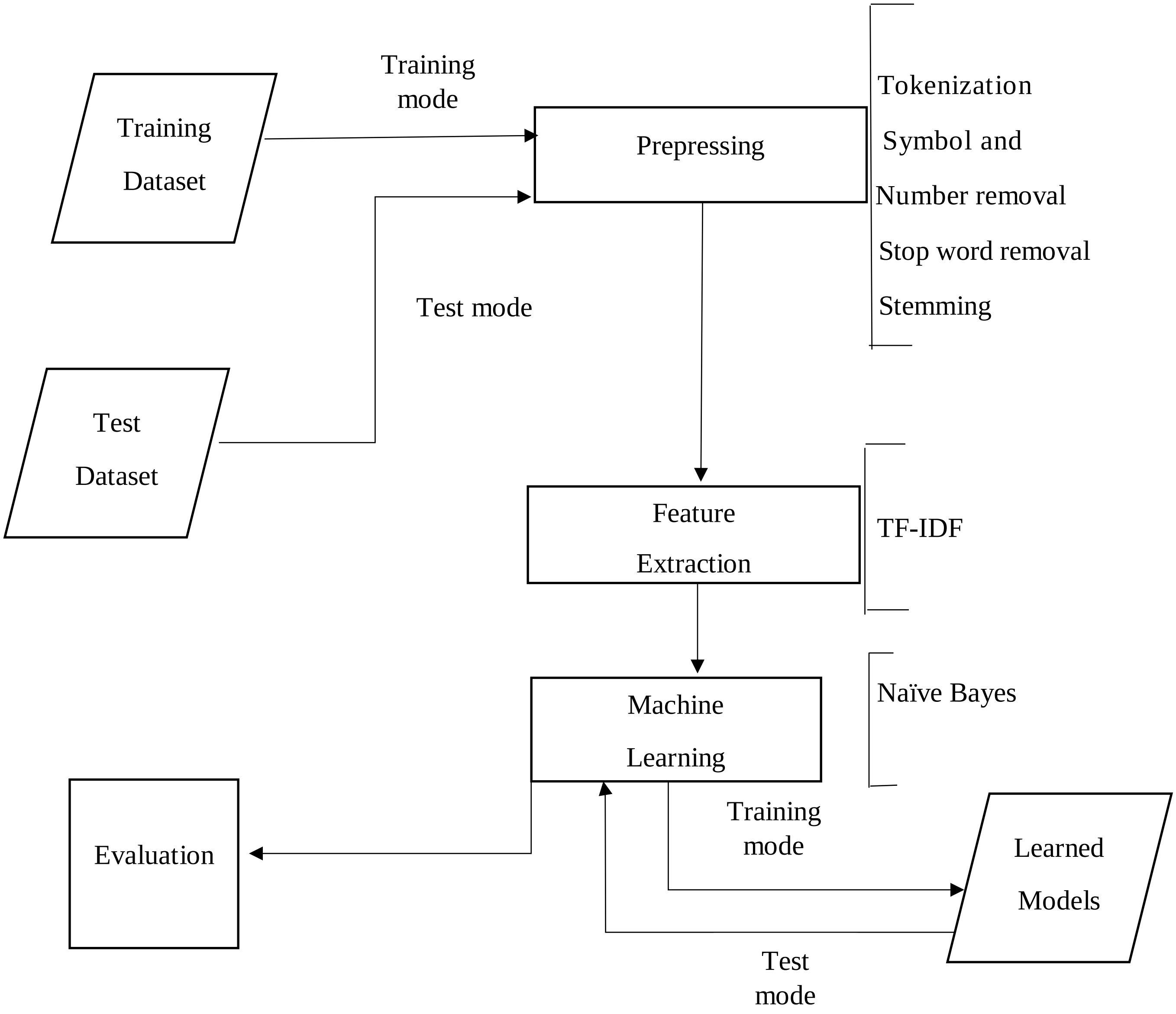


Figure 8:News classification system pipeline

3.2 Algorithms Studied and Implemented

3.2.1 Naïve Bayes Algorithm

Naive Bayes is a probabilistic machine learning algorithm that can be used in a wide variety of classification tasks. Typical applications include filtering spam, classifying documents, sentiment prediction etc. The name naïve is used because it assumes the features that go into the model is independent of each other. That is changing the value of one feature, does not directly influence or change the value of any of the other features used in the algorithm.

Since it is a probabilistic model, the algorithm can be coded up easily and the predictions made really quick. Real-time quick. Because of this, it is easily scalable and is traditionally the algorithm of choice for real-world applications that are required to respond to user's requests instantaneously.

Cons of Naïve Bayes:

If categorical variable has a category (in test data set) which was not observed in training data set, then model will assign 0(zero) probability and will be unable to make a prediction. This is often known as 'Zero Frequency'.

```
# Bernoulli Naive Bayes Algorithm
clf3 = Pipeline([
    ('vect', tfidfVectorizer),
    ('clf', BernoulliNB(alpha=0.01))
])
```

Figure 9:Implementation of Naive Bayes Algorithm

3.2.2 Suitability of Algorithms

The Naïve Bayes algorithm is used for Nepali News Classification because of the following reasons.

- Naive Bayes is an eager learning classifier and it is sure fast. Thus, it could be used for making predictions in real time.
- This algorithm is also well known for multi class prediction feature. Here we can predict the probability of multiple classes of target variable.
- Naive Bayes classifiers mostly used in text classification (due to better result in multi class problems and independence rule) have higher success rate as compared to other algorithms.

3.3 System Design

The following figures defines the architecture and the system design for this project.

1.3.1 Flow Chart

+

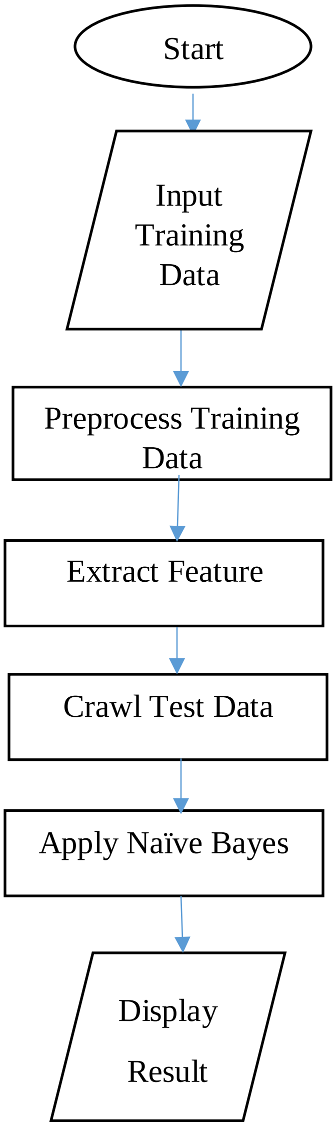


Figure 10: Flow Diagram of Overall System

Fig 10 shows the flow diagram of overall system. First, we input the training data, then the training data are preprocessed which includes symbol and number removal, stop word removal and word stemming. Then the features are extracted and the data are crawled. We apply the Naïve Bayes algorithm from which we classify the Nepali News into different categories.

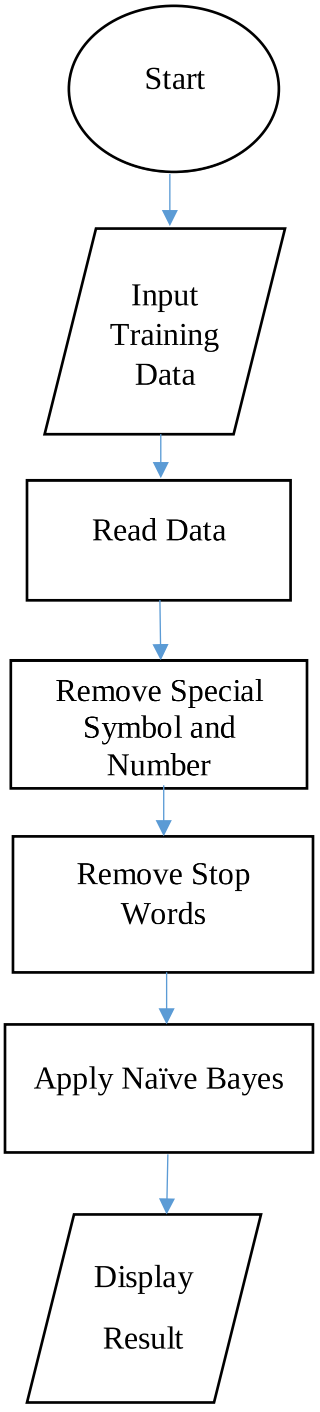


Figure 11: Preprocessing System

Fig 11 shows the preprocessing system which includes special symbol and number removal, stop words removal and apply the Naïve Bayes which the classifies the news.

3.3.3. Sequence Diagram

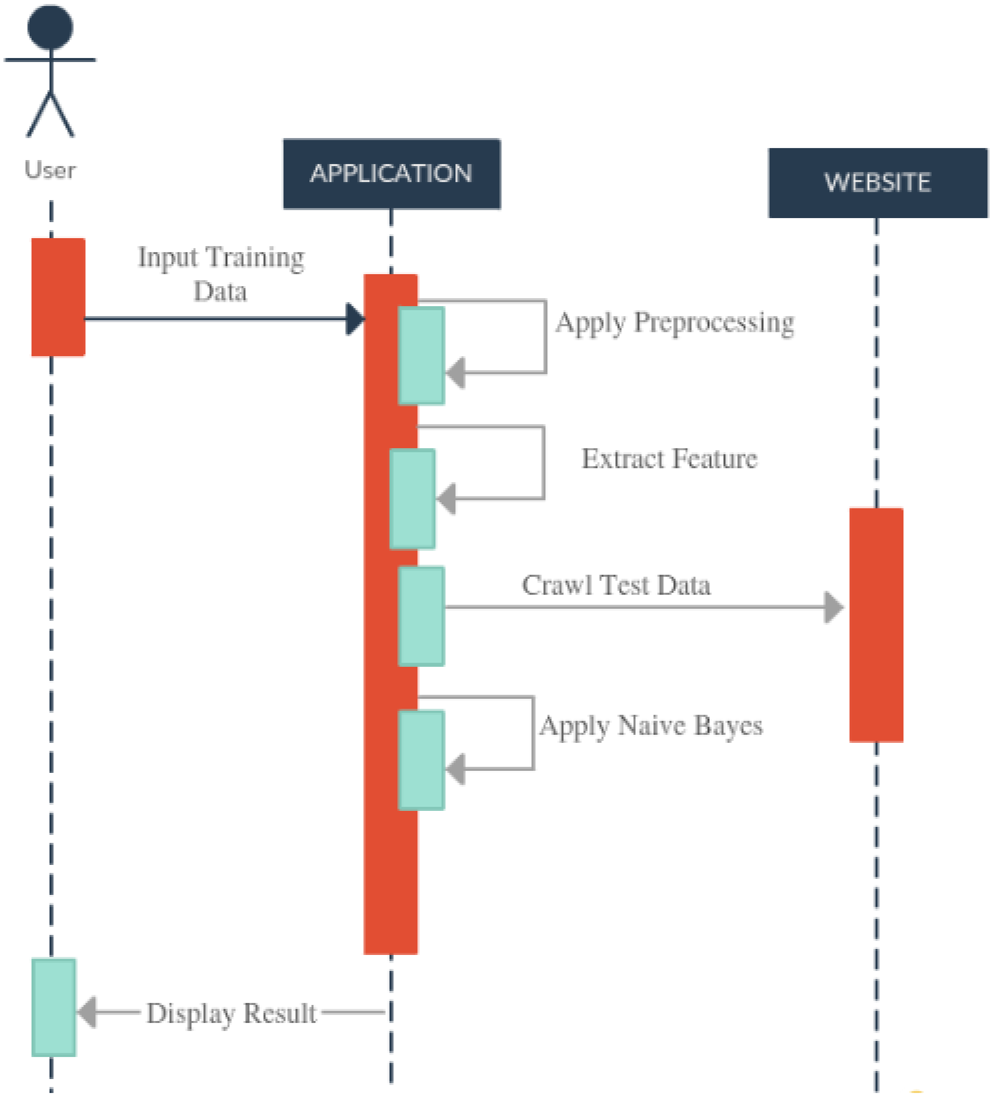


Figure 12: Sequence Diagram

Fig 12 shows the sequence diagram which shows the interaction between user, application and website. The user inputs the training data. We apply different preprocessing techniques, extracts feature and apply Naïve Bayes in the application and the content of the news is crawled from the website.

CHAPTER 4: IMPLEMENTATION AND EVALUATION

4.1 Tools and Technologies Used

Client side:

a) **HTML**: HTML is the standard markup language for documents designed to be displayed in a web browser. Web browsers receive HTML documents from a web server or from local storage and render the documents into multimedia web pages. HTML describes the structure of a web page semantically . Default characteristics for every item of HTML markup are defined in the browser, and these characteristics can be altered or enhanced by the web page designer's additional use of CSS.

b) **Bootstrap CSS**: Bootstrap is a free front-end framework for faster and easier web development. It includes HTML and CSS based design templates for typography, forms, buttons, tables, navigation, modals, image carousels and many other.

Server side:

a) **Python:** Python is a programming language which can be used on a server to create web applications. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object-oriented and functional programming. It also features dynamic name resolution, which binds method and variable names during program execution.

b) **Flask**: Flask is a microweb framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extension that can add application features as if they were implemented in Flask itself.

Nepali News Classification uses Flask framework which is a python-based framework. JavaScript, CSS and HTML pages has been used for development of front end while Python is used in backend.

4.2 Implementation

The major classes of the application are:

4.2.1 Classify data Class

When the classifier/user clicks the classify button in the browser, this is the first class that runs. we take the content of the news from the website, so this is used to crawl the content of the news from the website that is used for classification.

Input: It takes the content of the news from the user.

Process: It then calls the next class that is preprocessing class.

Output: It provides next class with the content of the news.

4.2.2 Preprocessing Class

All the preprocessing process such as removing special symbols, stop words, numbers and word stemming should be done. So, this class is used for all the preprocessing of the news that is extracted from the above class.

Input: This class takes raw news data as input.

Process: It calls functions namely, remove_numbers(), remove_stopwords()

Output: It produces processed text data that will act as input for next stage.

4.2.3 File: app.py

This is the main class, which is run first after the application is loaded in the browser. This class is responsible to trigger functions within the related classes. It is used to input the news content from the user.

4.2.4 Naïve Bayes Class

This class is used to implement the Naïve Bayes algorithm. Pickle a feature of python has been used to create a trained algorithm module which in runtime provides with the training data.

Input: This class takes extracted feature as input.

Process: It call function Pickle () and Predict () function.

Output: This class provides the output as classification of news.

4.3 Testing and their Results

During testing, content of the news from the testing datasets were used. For the testing purpose following 3 approaches were developed. Testing was conducted using three test cases:

Test Case 1: Cross validating the trained data

Test Case 2: Accuracy of the system

Test Case 3: Overall system testing

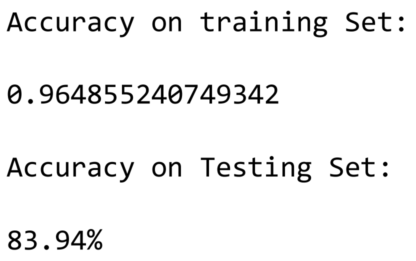
4.3.1 Cross validating the trained data

In this testing, any manual error was checked for the trained data set that was prepared. The entire dataset was cross checked since the weight has been assigned to the content of the news.

4.3.2 Accuracy of the system

```
('No of Samples, No. of Features ', (12918, 334041))

Bernoulli Naive Bayes
```



```
Accuracy on training Set:

0.964855240749342

Accuracy on Testing Set:

83.94%
```

Figure 13: Accuracy of the system

Table 3:Test Case 2

TC01 – Accuracy of the system
Precondition: Test and Trained data are given to the system.
Assumption: Weight is correctly assigned to all the data.
Test Steps: 1. Content of trained data was classified in regard to the preprocessing 2. Classification of the trained data and predicted classification is matched. 3. Accuracy is calculated by the mean of predicted and obtained classification.
Expected Result: The result must classify the news into predefined categories.
Generated Result: An accuracy of 83.94% was obtained.

4.3.3 Overall system testing

For overall system testing, 300 number of contents were fed in the system. The result obtained from the test is shown below in table 3.

Table 4: Test Case 3

Correct Analysis	Incorrect Analysis
268	32

CHAPTER 5: CONCLUSIONS AND LIMITATIONS

5.1. Conclusion

Nepali News Classification is successfully implemented using Flask framework of Python. A system has been built for categorizing the content of the news into different categories using the news article from a major Nepali language newspaper published in Nepal. This project evaluates some widely used machine learning techniques Naïve Bayes for automatic Nepali News classification. It classifies the news by analyzing the content of the news. Since the Nepali language is morphologically rich and complex, the text classifier needs to consider the specific language features before classifying the text. A self-created Nepali News Corpus with 16 different categories and total 16719 documents collected by crawling different online national news portal is used. Therefore, Nepali News Classification was developed using Naïve Bayes. Hence, further analysis is required to determine a concrete conclusion.

5.2. Limitations

- Many defining words can result in less accuracy in the application.
- The input can only be in text format.
- Categorization on basis of different variant section news of different newspaper makes classification difficult.
- Takes time for the output, excessive delay was not encountered.
- Actual meaning of the sentence is sometimes misinterpreted affecting the application's accuracy.

REFERENCES

[1] M. Puteh, "Sentiment Mining of Malay Newspaper (SAMNews) Using Artificial Immune System", in *WCE 2013*, London,UK, 2013.

[2] T. Shahi, "Nepali News Classification using Naive Bayes, Support Vector Machine and Neural Networks", 2018. [Online]. Available: https://www.researchgate.net/publication/324098346_Nepali_news_classification_using_Naive_Bayes_Support_Vector_Machines_and_Neural_Networks.

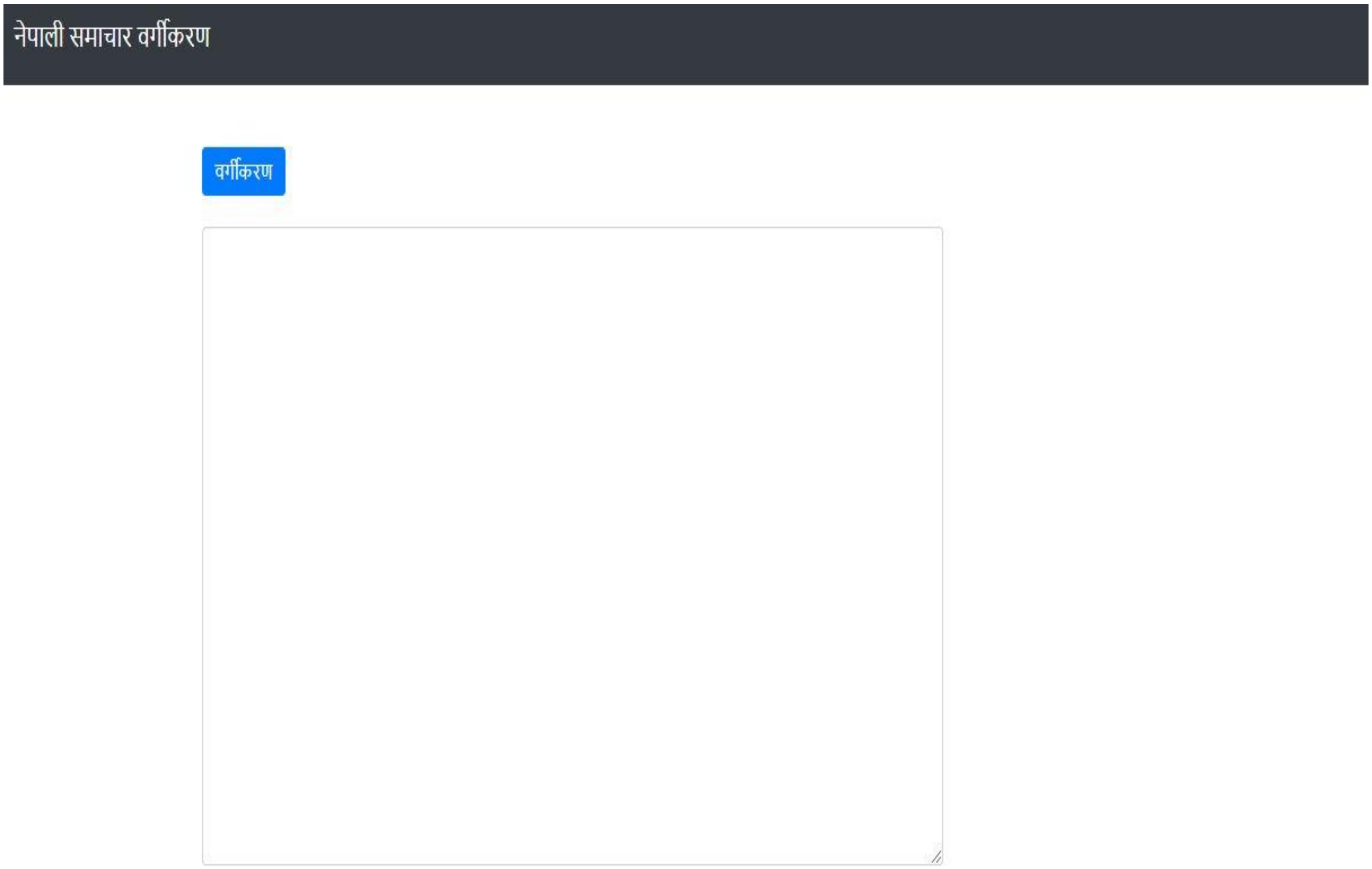
[3] S. Bhandari, "Predicting Sentiment on News Data", How Sentiment Analysis incorporates power of NLP and Text Analysis, 2018.

[4] G. Kaur, "News Classification and Its Techniques: A Review", Iosrjournals.org, 2016. [Online]. Available: http://www.iosrjournals.org/iosr-jce/papers/Vol18-issue1/Version-3/D018132226.pdf.

[5] D. Dongol, "Automated News Classification using N-gram Model and Key Features of Nepali Language", 2018.

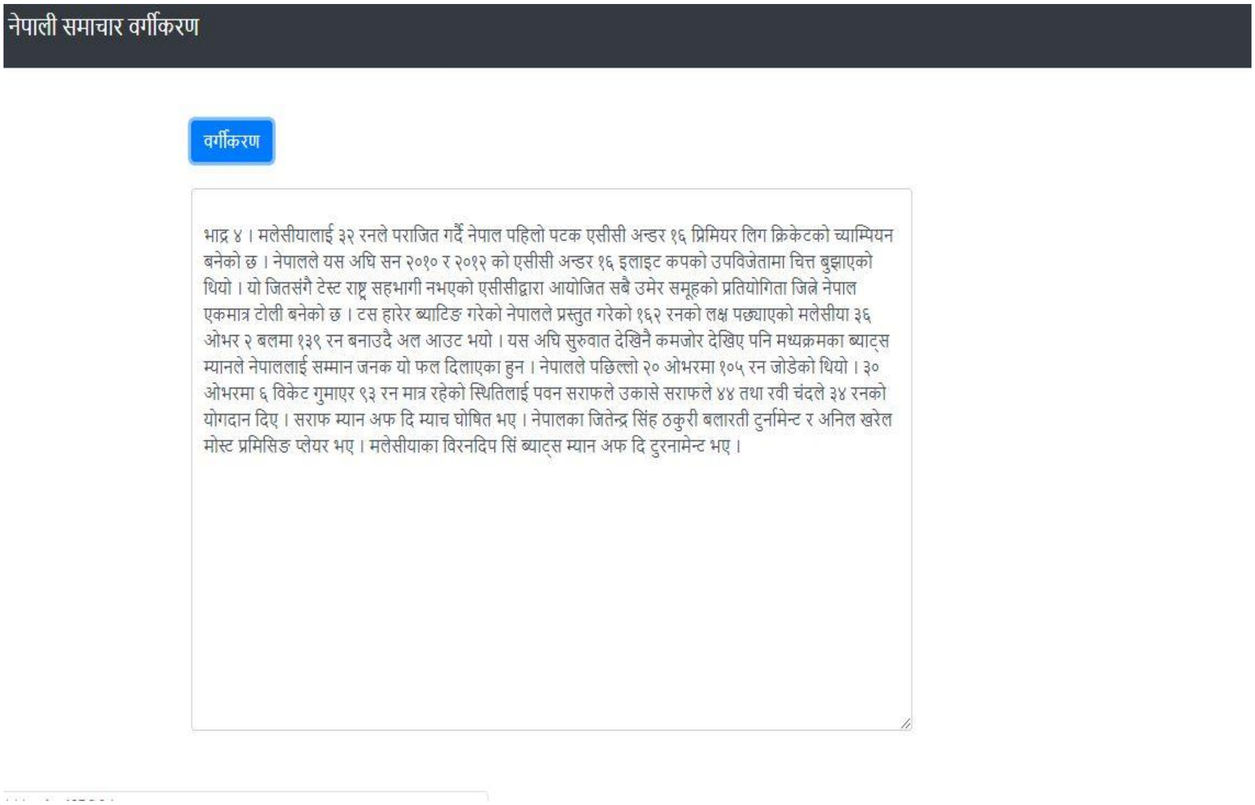
[6] N. Naw, "Twitter Sentiment Analysis Using Support Vector Machine and K-NN Classifiers", International Journal of Scientific and Research Publications (IJSRP), vol. 8, no. 10, 2018. Available: 10.29322/ijsrp.8.10. 2018.p8252.

APPENDIX I



Screenshot of Landing Page

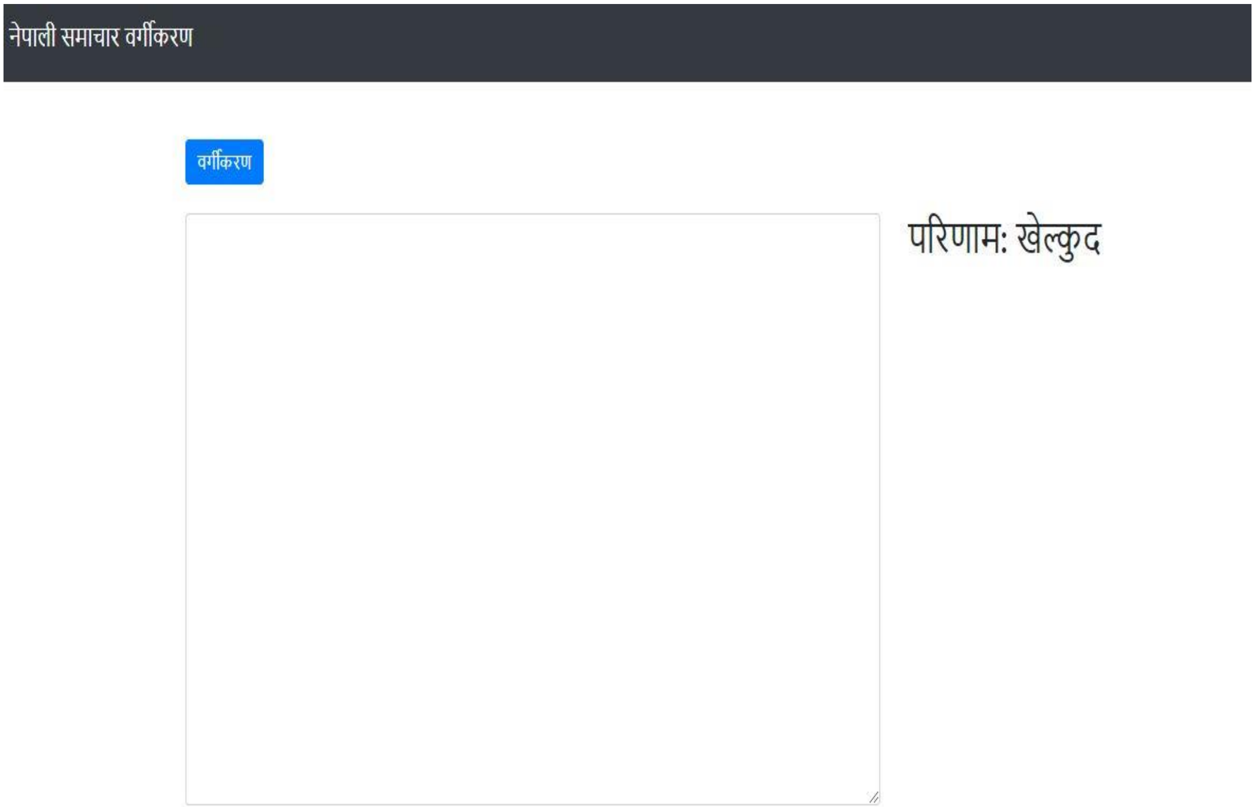
APPENDIX II



भाद्र ४ । मलेसीयालाई ३२ रनले पराजित गर्दै नेपाल पहिलो पटक एसीसी अन्डर १६ प्रिमियर लिग क्रिकेटको च्याम्पियन बनेको छ । नेपालले यस अघि सन २०१० र २०१२ को एसीसी अन्डर १६ इलाइट कपको उपविजेतामा चित्त बुझाएको थियो । यो जितसंगै टेस्ट राष्ट्र सहभागी नभएको एसीसीद्वारा आयोजित सबै उमेर समूहको प्रतियोगिता जित्ने नेपाल एकमात्र टोली बनेको छ । टस हारेर ब्याटिङ गरेको नेपालले प्रस्तुत गरेको १६२ रनको लक्ष पछ्याएको मलेसीया ३६ ओभर २ बलमा १३९ रन बनाउदै अल आउट भयो । यस अघि सुरुवात देखिनै कमजोर देखिए पनि मध्यक्रमका ब्याट्स म्यानले नेपाललाई सम्मान जनक यो फल दिलाएका हुन । नेपालले पछिल्लो २० ओभरमा १०५ रन जोडेको थियो । ३० ओभरमा ६ विकेट गुमाएर ९३ रन मात्र रहेको स्थितिलाई पवन सराफले उकासे सराफले ४४ तथा रवी चंदले ३४ रनको योगदान दिए । सराफ म्यान अफ दि म्याच घोषित भए । नेपालका जितेन्द्र सिंह ठकुरी बलारती टुर्नामिन्ट र अनिल खरेल मोस्ट प्रमिसिङ प्लेयर भए । मलेसीयाका विरनदिप सिं ब्याट्स म्यान अफ दि टुरनामेन्ट भए ।

Screenshot of News Page

APPENDIX III



परिणाम: खेल्कुद

Screenshot of classification of the news page