EVEREST ENGINEERING COLLEGE

(Affiliated to POKHARA UNIVERSITY)

A Minor Project Proposal

On

NEPALI FAKE NEWS DETECTION

Submitted By:

[Abhishek Sah]	[21075393]
[Abhishek Shrestha]	[21075394]
[Dibas Timilsena]	[21075410]
[Nirmal Khatri]	[21075419]

Submitted To:

Department of Computer and IT Engineering

Everest Engineering College

Sanepa-2, Lalitpur

DEC 27, 2024

ACKNOWLEDGEMENT

We would like to extend our sincere and heart felt gratitude to Er. Narayan Sapkota and Er. Mukunda sir for their invaluable assistance and providing expert guidance throughout the course of this project. We would also like to express our profound appreciation to the Department of Computer and IT Engineering of our college.

We would be grateful for any suggestions or feedback that could contribute to the further enhancement of this project.

Abhishek Sah

Abhishek Shrestha

Dibas Timalsena

Nirmal Khatri

ABSTRACT

In recent years, due to the booming development of online social networks, fake news for various commercial and political purposes has been appearing in large numbers and widespread in the online world. With deceptive words, online social network users can get infected by online fake news easily, which has brought about tremendous effects on the offline society already. An important goal in improving the trustworthiness of information in online social networks is to identify the fake news timely. The purpose of this project is to outline the creation of an algorithm for detecting fake news articles. Through this project it will provide users with the ability to detect whether the news they are being provided is authentic News or not.

We will be performing binary classification of various news articles available online with the help of public datasets like Kaggle datasets. We are going to use python as programming languages and logistic regression machine learning algorithm Concepts pertaining to Artificial Intelligence, Natural Language Processing (NLP) and Machine Learning models provide the ability to classify the Nepali news as fake or real.

Keywords: *Machine Learning, dataset, model, Natural Language processing*

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

CFG:	Context Free Grammer
IDE:	Integrated Development Environment
NLP:	Natural Language Processing
NLTK:	Natural Language Toolkit
TF IDF:	Term Frequency - Inverse Document Frequency
UI:	User Interface
CSV:	Comma Separated Value

Chapter 1: INTRODUCTION

1.1 Background

These days fake news is creating different issues from sarcastic articles to a fabricated news and plan government propaganda in some outlets. Fake news and lack of trust in the media are growing problems with huge ramifications in our society.

Obviously, a purposely misleading story is fake news but lately blathering social media's discourse is changing its definition. Some of them now use the term to dismiss the facts counter to their preferred viewpoints. With the current usage of social media platforms, consumers are creating and sharing more information than ever before, some of which are misleading with no relevance to reality. Given a multi-source news dataset and social contexts of news consumers (social media users), the task of fake news detection is to determine if a news item is fake or real

Machine Learning is a way for computers to learn from data without being explicitly programmed to do so. Instead of being given exact instructions, the computer is given lots of examples and it figures out patterns on its own. Think of it like teaching a child by showing them different objects; over time, they learn to recognize new objects without needing direct instruction. In machine learning, this ability to detect patterns and improve from experience helps the computer make predictions or decisions without human intervention.

1.2 Problem Statement

Primarily, political sectors are the main targets for fake news. but it is not limited to this. Lately, with the outbreak of the COVID-19 pandemic, lots of bogus news and myths regarding the disease have gone viral on the Internet. This has affected the mental well-being

People are often deceived by the fake news circulating on the Internet mainly due to three reasons:

- First of the people during this difficult time, the information confirming their preexisting attitudes is preferred (selective exposure).
- Second, the information consistent with their preexisting beliefs is more persuasive (confirmation bias).
- Third, people are more inclined to accept the information that pleases them (desirability bias).

1.3 Objectives

The objective of this project is to address the spread of fake news by applying machine learning algorithms with (NLP) to various datasets. Various (NLP) approaches can be used to analyze the content and style of the news to detect the context and facts in the article.

1.4 Scope and Application

The scopes of this proposed project are:

- Enables faster identification of fake news, helping users and stakeholders act quickly to counter misinformation.
- Helps prevent the spread of fake news during emergencies or crises, reducing panic and misinformation-driven actions.
- Prevents financial scams and misinformation about businesses products, protecting consumers and maintaining economic trust.

The application of this proposed project are:

- Useful for news agencies to verify the authenticity of news articles before publication.
- Assists in monitoring and regulating fake news to maintain social harmony and prevents misinformation driven
- Offers a user-friendly platform for individuals to check the authenticity of news they encounter online.

Chapter 2: LITERATURE REVIEW

In the research made by **Soniya C. J And Shrihari M.** where they have mentioned about the importance of **fake news detection** as users can easily find whether the news is fake or not. Here they used Machine learning Algorithms for classifications likes: Naïve Bayes Classifier, Random Forest, Logistic regression, Passive Aggressive Classifier. The primary aim of the research is to identify patterns in text that differentiate fake articles from true news.[1]

In the journal of student research **Qiheng. G and Nicle. L** Where they built a project to detect fake news using machine learning algorithmby using NLP(natural language processing) to interpret and sort words and machine learning techniques such as SVM and gradient boosting to differentiate fake news from realnews. [2]

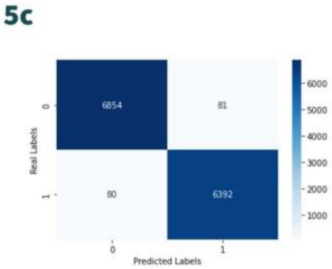


Figure 1:Confusion Matrices

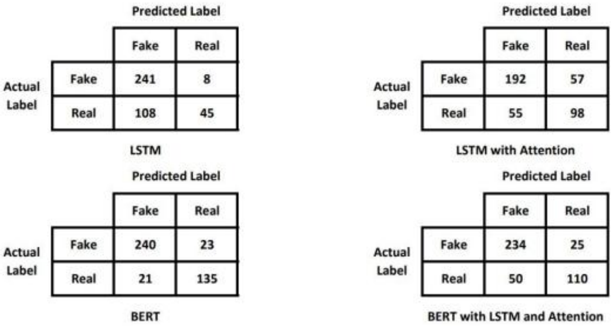
In the research made by **Hadeer. A, Issa. T and Sherif. S**, entitled "Detection of online fake news using n-gram analysis and machine learning techniques" where they have presented an alternative to detect fake news. Here, they have worked on concepts and algorithm of machine learning like

*Bag of Words (BOW by using system (NLP)natural language processing to interpret and sort old and machine learning techniques such as sbm and gradient boosting to differentiate fake news from real news) [3]

* Logistic regression

* Vector Machine

Research made by Neha Sake and Prakash Paudal they built a model which detect the fake news using deep neural networks Department of computer science and Engineering Kathmandu University in Dhulikhel NepalThey use LSTM,BERT with algorithm



		Predicted Label	
		Fake	Real
Actual Label	Fake	241	8
	Real	108	45

LSTM

		Predicted Label	
		Fake	Real
Actual Label	Fake	192	57
	Real	55	98

LSTM with Attention

		Predicted Label	
		Fake	Real
Actual Label	Fake	240	23
	Real	21	135

BERT

		Predicted Label	
		Fake	Real
Actual Label	Fake	234	25
	Real	50	110

BERT with LSTM and Attention

Figure 2: Confusion matrix of result opened by different model

In the paper by **Shaina. R** and **Chen.D** of Fake news detection based on news content and social contexts: a transformer-based refers to the transformer architecture, which facilitates representation learning from fake news data and helped them detect fake news early. The input is taken as the news items, social context and associated side information. The output is based on one of two labels 'false' or 'real'.

The highest accuracy for the proposed FND-NS MODEL is 74.8%, whereas precision and recall are 72.4% and 77.6% respectively. Area Under the Curve is measured as 70.4% and average precision among all models is 71%. [4]

Chapter 3: METHODOLOGY

3.1 Workflow Diagram

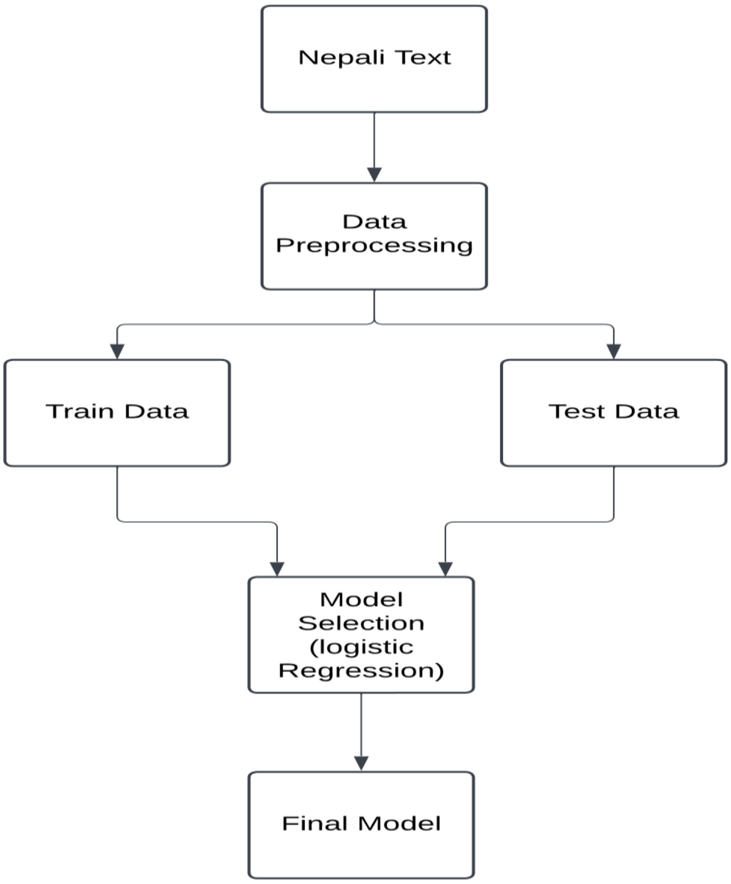


Figure 3:Workflow Diagram

3.2 Working Principle

3.2.1 Data Description

A dataset is a collection of data used for training and testing a machine learning model. It is a fundamental component, as the model learns patterns and relationships in the data to make predictions or classifications. So we find the label datasets in GitHub.

3.2.2 Data Preprocessing

Good data is essential for creating clear visualizations and accurate machine learning models. Preprocessing cleans and organizes the data, making it easier to work with and helping machine learning algorithms perform better. The preprocessing techniques include stop words removal, symbol and number removal tokenization and stemming.

Tokenization

Tokenization is the process of breaking down text into smaller units called tokens. These tokens can be words, phrases, or symbols, and tokenization is a crucial step in natural language processing (NLP) and machine learning (ML) for our project.

Lemmatization and steaming

Stemming involves cutting off prefixes or suffixes from a word to obtain its root form. This process is often crude and may not produce a valid word. Stemming is a technique that reduces a word to its base word, called stem, aiding the process of text processing. For example, “घरमा” becomes “घर”, “देशको” becomes “देश”, “मलाई” becomes “म” after stemming.

Stop word removal

Stop words in documents are the words which occur frequently that may or may not have any meaningful uses for information retrieval process. These are the common words. It includes language specific determiners, conjunctions, and postpositions. The stop words list for English and other language are easily available but there is not any standard stop words list for the Nepali language.

Original Sentence: "नेपालमा कोरोना भाइरसको सङ्क्रमण बढ्दैगएकोछ।"

Sentence with Stop Words Removed: "नेपाल कोरोना भाइरस सङ्क्रमण बढ्दै गएको।"

Special symbol and number removal:

Special symbols and numbers, those do not have much importance in classification, are removed. The punctuation in the text consists of different types of symbols. Some of symbols used in Nepali text are given below.

Symbols:,) (! : - / ? ।

Numbers: ० १ २ ३ ४ ५ ६ ७ ८ ९

TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF is a widely used feature vector representation technique for the text analysis in natural language processing. It is a statistical method to find the importance of words in a document. Due to complex word segmentation of Nepali language, TF-IDF is one of the mostly used, easy methods to extracts features from text. It mainly consists of two parts.

Term Frequency (TF): TF represents occurrence of terms in a document. In TF, scoring is given to words based on the frequency. The frequency of words is dependent on the length of the document, i.e. in large size document, word occurs more as compared to small size documents. The TF can be calculated as;

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t'd}}$$

Where:

- $f_{t,d}$: the number of times t appears in document d.
- N_d: the number of terms in document d.

Inverse Document Frequency (IDF): It is the number of documents that contain a term in the collection of documents. It is a document-level statistic that gives a score on the basis of document level. The scoring is given to a word based on how a word is rare across all documents. The IDF of a rare term is high, as compared to the IDF of a frequent term.

$$idf(t,D) = log \frac{N}{|\{d \in D: t \in d\}|}$$

Where:

- |D|: total number of documents.
- $\{d \in D: t \in d\}$: Numbers of documents containing the term t.

The TF- IDF formula is :

TF-IDF(t,d,D) = TF(t,d) * IDF(t,D)

3.2.3 Feature Engineering

Feature engineering plays a crucial role in improving the performance of machine learning models. It involves selecting the most relevant features (data attributes) that will contribute significantly to the model's predictive power, transforming raw data into a format that is better suited for the learning algorithm, and creating new features that can enhance model accuracy.

3.3 Machine learning Algorithm

A variety of machine learning have been employed to automatically classify news articles as fake or real. these models leverage text features linguistic patterns and sometimes metadata to distinguish between deceptive and factual news some of the models we will used in our projects are. Text preprocessing, Feature extraction. We can add more features according to our time schedule for better performance.

Logistic regression

Logistic Regression is a widely used machine learning algorithm, particularly suited for binary classification tasks, making it ideal for detecting fake news in Nepali text. The model predicts the probability that a given news article belongs to one of two categories—fake or real. It does this by analyzing the relationship between the input features (such as words, phrases, or sentence structures) and the log-odds of the article being fake or real.

The core of Logistic Regression lies in the logistic function (sigmoid), which maps the raw prediction scores to a probability between 0 and 1. This is especially useful in fake news detection as it allows for confidence levels in classifying news articles.

The logistic function is generally expressed as:

$$p(x) = \frac{1}{\{1 + e^{\{-(\beta_0 + \beta_1)\}}\}}$$

Where:

$\beta_0 = -\frac{\mu}{s}$: intercept or vertical intercept in the linear term.

β_1=1\s: inverse scale parameter.

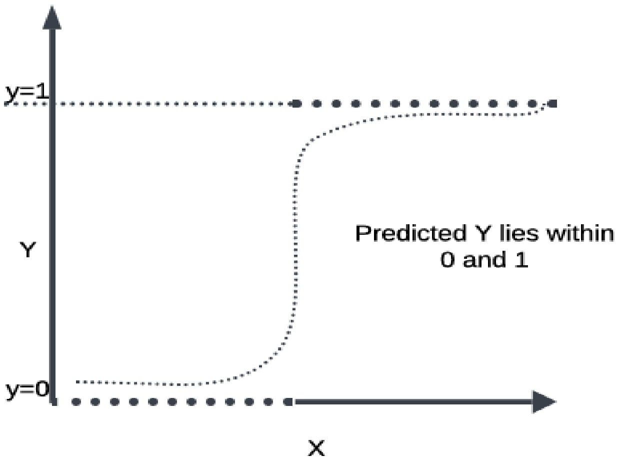


Figure 4: Logistic Regression

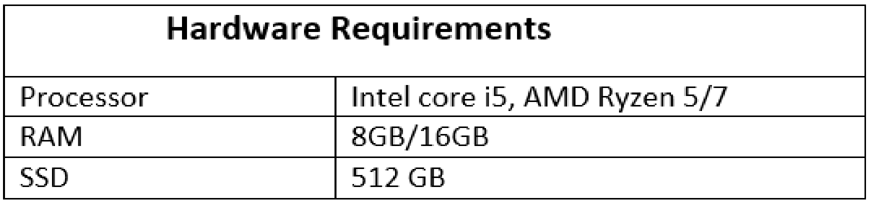
3.4 Hardware and Software Required

Hardware components required for our project are:

- PC/Laptop

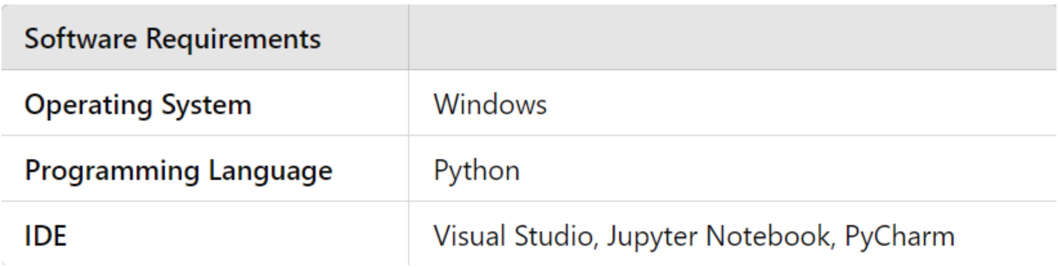
Minimum Requirement:

Table 1: Hardware Requirements



Hardware Requirements	
Processor	Intel core i5, AMD Ryzen 5/7
RAM	8GB/16GB
SSD	512 GB

Table 2: Software Requirements



Software Requirements	
Operating System	Windows
Programming Language	Python
IDE	Visual Studio, Jupyter Notebook, PyCharm

3.5 Tools and Libraries

Table 3: Tools and Library

Technology/Library	Description
HTML (Hypertext Markup Language)	Used to create the structure of web pages. Without HTML, a web page would be a jumbled mess of text and images.
CSS (Cascading Style Sheets)	Used to style and layout the web page, making it stylish and attractive. Without CSS, the webpage would be plain text on a white background.
JavaScript	Adds interactivity and features to improve user experience and enable the webpage to execute actions.
NLTK (Natural Language Toolkit)	Provides tools for tokenization, stop word removal, stemming, lemmatization, and more for text processing.
Python	A high-level, general-purpose programming language emphasizing code readability through significant indentation. It is dynamically typed and garbage collected.
Jupyter Notebook	An interactive environment to write and execute code in real-time. It supports various programming languages, with Python being the most commonly used.
Scikit-Learn	Often used for converting text into numerical features using techniques like Bag of Words, TF-IDF, and handling tokenization for machine learning models.
Pandas	A Python library for data manipulation and analysis, offering data structures and operations for numerical tables and time series.
NumPy	Supports fast and efficient matrix operations, dot products, and other linear algebra functions, essential for machine learning models.
Matplotlib	A powerful library for creating visualizations in Python, crucial for analyzing and presenting data in the project.

3.6 Hyperparameter Tuning

In the context of fake news detection, hyperparameter tuning is essential for ensuring that the model performs optimally on the task of identifying whether an article is true or false. The goal is to find the combination of hyperparameters that leads to the best model performance (accuracy, precision, recall, F1-score, etc.)

Efficient methods for hyperparameter tunning are:

Random Search

Instead of trying every combination, random search samples random combinations of hyperparameters. This is often more efficient than grid search and can give good results with less computational cost.

3.7 Model Evaluation Metrics

The precision, recall, and F1-scores for our project are as follows:

Class 0:

- Precision: 0.93
- Recall: 0.91
- F1-score: 0.92

Class 1:

- Precision: 0.91
- Recall: 0.93
- F1-score: 0.92

Overall (Macro average):

- Precision: 0.92
- Recall: 0.92
- F1-score: 0.92

Overall (Weighted average):

- Precision: 0.92
- Recall: 0.92
- F1-score: 0.92

3.8 Data Collection

After some research our team found that in several sources, we found the news detection data sets often are available for free or via academic. We were able to find Nepali fake news datasets in (CSV) format on GitHub, which we are using for our projects.

Chapter 4: Implementation

4.1 Problem Faced

While building this project we had encounter the error such as:

- Lack of proper stemming library for Nepali words: During the development of the Nepali fake news detection system, we faced the challenge of not having an established library for stemming Nepali word. Existing NLP libraries and frameworks primarily cater to widely spoken languages like English, making it difficult to preprocess Nepali text.
- Challenge in Stop Words Removal: While building the Nepali fake news detection system, we encountered the issue of a lack of pre-defined and reliable stop word lists for the Nepali language. To overcome this limitation, we manually curated a list of Nepali stop words by analyzing text corpora and identifying frequently occurring but contextually insignificant words.

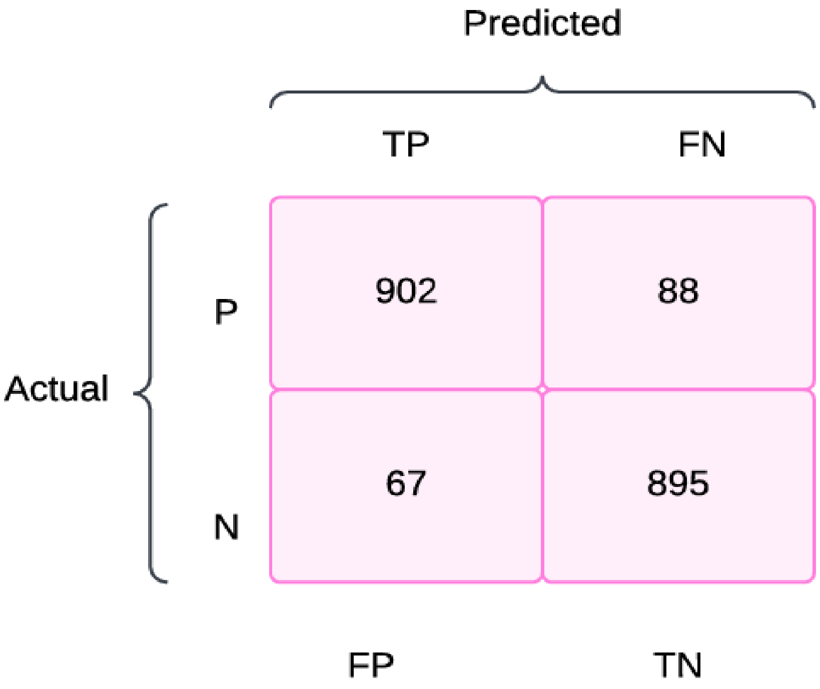
4.2 worked completed

- **Data preprocessing and cleaning**: We successfully implemented a Nepali text preprocessing module, including tokenization, punctuation handling, lowercase conversion, and removing manually curated stop words, ensuring structured and noise-free data for analysis.
- **Model Development and Testing:** The fake news detection module has been trained and tested using supervised machine learning techniques. We utilized labeled datasets of Nepali news articles to fine-tune the model for accurate classification of fake and genuine news.

The current status of our project is progressing well toward detecting fake news effectively. The module demonstrates reliable classification with tailored preprocessing for Nepali text, though some preprocessing tasks remain to be completed before full deployment.

4.3 Performance metrics

The confusion metrics of our project is:



	TP	FN
P	902	88
N	67	895
	FP	TN

Figure 5: Confusion metrics

4.4 Remaining Works

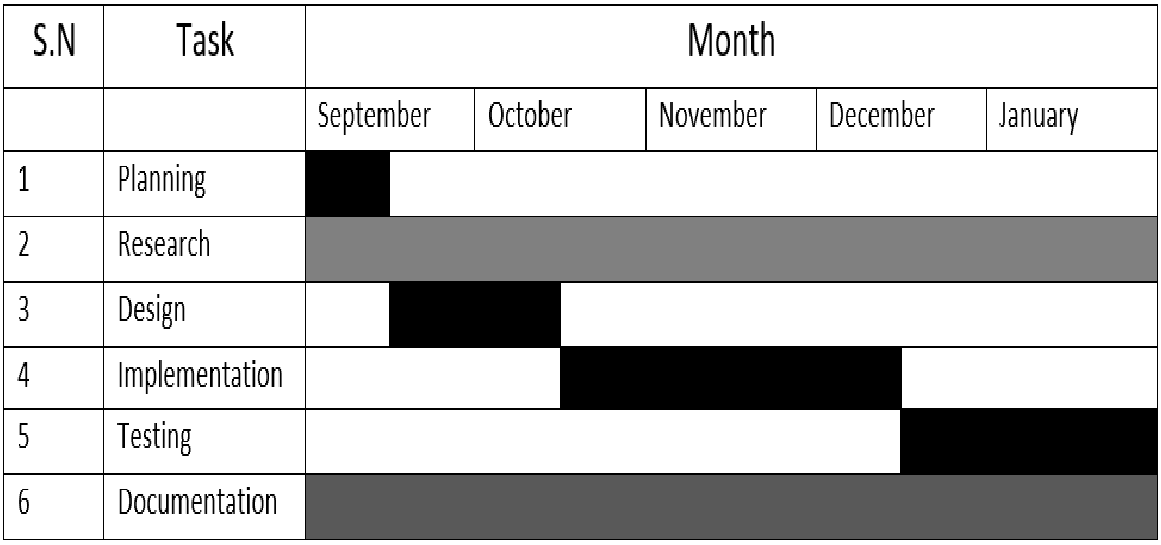
- **Stemming**: The text preprocessing pipeline requires the implementation of a stemming module to further refine Nepali text. This step is essential to ensure words are normalized to their root forms, enhancing the accuracy of feature extraction and model predictions.
- **User Interface (UI) Development**: The development of a user-friendly UI for the fake news detection system is still in progress. The UI will allow users to input Nepali text or news articles, providing seamless interaction and displaying results effectively.

Chapter 5: PROJECT SCHEDULE

5.1 Schedule

The feasibility of this project in terms of time. The estimated duration of our project is given with the performed activities in the following given char

Table 6: Gantt chart



S.N	Task	Month				
		September	October	November	December	January
1	Planning	■				
2	Research	▨	▨	▨	▨	▨
3	Design	■	■			
4	Implementation		■	■	■	
5	Testing				■	■
6	Documentation	▨	▨	▨	▨	▨

Chapter 6: FEASIBILITY ANALYSIS

We have outlined the project's feasibility, covering its schedule, budget, technical requirements, and operational aspects. This comprehensive analysis will serve as a valuable reference for decision-making and guiding future development efforts.

6.1 Financial Feasibility

Since our project is software based, and we do not need any sort of hardware components, the cost estimation for our project management is Nrs.2000.

6.2 Technical Feasibility

The project will use jupyter notebook visual studio code as ide which are open source and readily available throughout the internet. So, project is technically feasible.

6.3 Time Feasibility

The planning phase was completed by mid-September, and immediately afterward, the research phase began. The research phase is expected to be completed by October 1st. Once the research is done, the design phase will start and continue until mid-October. After the design phase concludes, the coding phase will begin. Testing will take place from mid-December until the project wraps up. Documentation started in mid-September and will continue throughout the remainder of the project until its completion.

Chapter 7: CONCLUSION

We are currently building a Nepali fake news detection system using logistic regression, where the model classifies news articles as real or fake based on the dataset. The user interface (UI) for this system is in the development phase, and once completed, it will allow users to easily input news articles and receive predictions on their authenticity. The model aims to achieve high accuracy in detecting fake news tailored to the Nepali language.

References

[1] S. M. R. Sonia C J, "Fake news detection," *VISVESVARAYA TECHNOLOGICAL UNIVERSITY,* no. 1, pp. 1-26, 2021-2022.

[2] Q. a. N. .L, "Using Machine Learning Algorithms to Detect Fake," *high school edition,* vol. 11, no. 4, p. 9, 2022.

[3] h. a. i. traore, "Detection of Online Fake News Using N-Gram," *ECE Department, University of Victoria,* p. 13, 2017.

[4] S. .. a. C. .D, "Fake news detection based on news content and social contexts:," *Ryerson University,* p. 28, 2022.

[5] N. S. a. P. Paudel, "Detection of fake news using deep neural networks," *kathmandu university,* p. 7, 2022.

[6] B. K. bal, "A Nepali Rule Based Stemmer and its performance on different NLP applications," 2004.

[7] A. sakya, "A Nepali Rule Based Stemmer and its performance on different NLP applications," 2020.