EVEREST ENGINEERING COLLEGE

(Affiliated To POKHARA UNIVERSITY)

Minor Project FINAL REPORT

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

FEB 25, 2025

ACKNOWLEDGEMENT

We would like to extend our sincere and heart felt gratitude to Er. Narayan Sapkota and Er. Mukunda Poudel 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 Timilsena

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
CSV:	Comma Separated Value
IDE:	Integrated Development Environment
NLP:	Natural Language Processing
NLTK:	Natural Language Toolkit
ROC:	Receiver-operating characteristic curve
TF IDF:	Term Frequency - Inverse Document Frequency
UI:	User Interface

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 are 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.

1.5 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

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.

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.

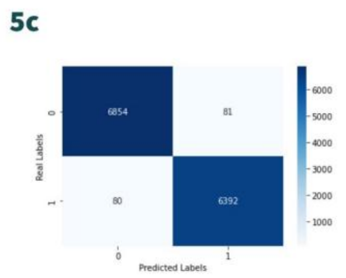

Figure 1: Confusion Matrix

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)

* 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

LSTM

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

LSTM with Attention

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

BERT

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

BERT with LSTM and Attention

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

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%.

Chapter 3: METHODOLOGY

The Nepali Fake News Detection System utilizes a supervised machine learning algorithm, specifically Logistic Regression, to classify news articles as either fake or real. This approach requires a labeled dataset. The detailed architecture of the system is outlined below.

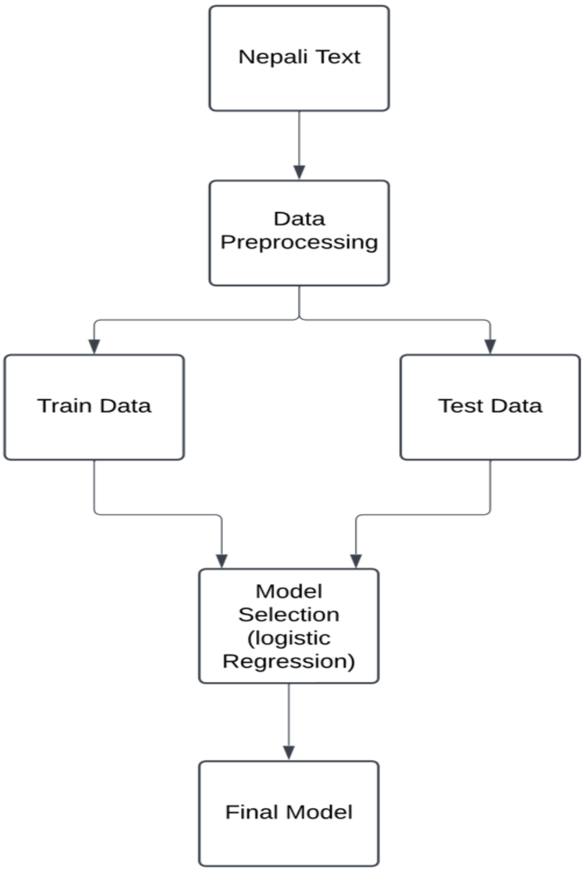

Figure 3: Workflow Diagram

3.1 Data Description

The dataset for the Nepali fake news detection project includes 10,000 true news articles and 9200 fake news articles. This balanced dataset allows the model to learn effectively, with equal representation of both categories. The data is sourced from GitHub and is enriched with news scraped from trusted Nepali outlets like Gorkha Patra, Shila Patra, Rato Pati, and Online Khabar, providing a wide range of topics and writing styles for training the model. The features include the text content, article headlines, and labels indicating whether the news is true or fake.

3.2Data 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.

Figure 4: Workflow for processing Nepali Text

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.

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 :

$$\text{TF-IDF}(t,d,D) = \text{TF}(t,d) * \text{IDF}(t,D)$$

3.3 Feature Engineering

In the Feature Engineering process for the Nepali fake news detection system, the text data was preprocessed by tokenizing, removing stopwords (using a custom Nepali stopword list), and applying stemming to reduce words to their root form. Features were then extracted using TF-IDF (Term Frequency-Inverse Document Frequency), which helped capture the importance of words relative to the entire corpus. To ensure the most relevant features were used, techniques like removing low-frequency words and analyzing feature correlation were applied.

3.4 Model Selection and Justification

Several machine learning models were evaluated for the Nepali Fake News Detection project, focusing on efficiency, performance, and adaptability. Among them, Naïve Bayes, Gradient Boosting Classifier, and Logistic Regression were considered. After comparison, Logistic Regression provided the best results.

Table 3: Comparative Analysis of Classification Algorithms

	Precision		Recall	F1 score	Testing Accuracy
Gradient Boosting	0	0.99	0.96	0.97	0.97
	1	0.96	0.99	0.97	
Logistic Regression	0	1.00	0.97	0.98	0.98
	1	0.97	1.00	0.98	
Naive Bayes	0	0.91	1.00	0.95	0.94
	1	1.0	0.89	0.94	

The table highlights Logistic Regression's strong performance with a 0.98 testing accuracy, perfect precision for class '0', and high recall for both classes. This, coupled with its high F1-score, confirms its effectiveness for classifying Nepali news, validating its selection due to its accuracy, simplicity, and efficiency, especially with sparse text data.

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.

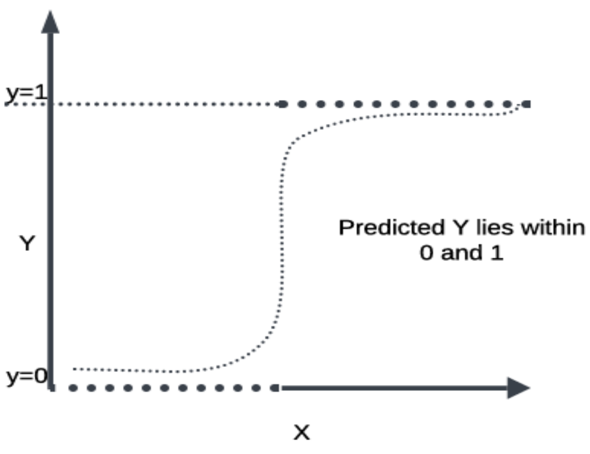

Figure 5 : Logistic Regression

3.5 Training and Testing Strategy

To ensure the effectiveness and generalization of the Nepali Fake News Detection Model, a structured training and testing strategy was implemented. The dataset was split into 90% training data and 10% testing data, ensuring that the model learns effectively from a larger portion of the data while reserving a small set for evaluation.

3.5.1 Overfitting Issue and Regularization Solution

During model evaluation, overfitting was observed, where the model performed well on training data but showed a drop in accuracy on unseen test data. To address this, L1 and L2 Regularization (Ridge Regression) was applied in Logistic Regression to control model complexity and prevent large coefficients from dominating predictions. This helped in:

- Reducing Variance: Preventing the model from memorizing training data patterns.
- Enhancing Generalization: Ensuring consistent performance on new, unseen data.

3.6 Hyper parameter Tunning

Hyperparameter tuning is a crucial step in machine learning to optimize model performance in this project we used grid search with cross -validation to find the best hyperparameters for the logistic regression model.

we applied both Grid Search and Random Search to tune the hyperparameters of the Logistic Regression model. After comparing the results, Grid Search provided the best performance. The final model achieved an accuracy of 99.13% on the test dataset, indicating highly accurate predictions. The classification report shows a precision, recall, and F1-score of 0.99 for both classes, meaning the model correctly identifies both fake and real news with minimal errors. This high performance suggests that the selected hyperparameters effectively optimized the model for detecting Nepali fake news

For tuning, we considered the following hyperparameters:

- **Regularization strength (C)**: [0.001, 0.01, 0.1, 1, 10, 100]
- **Regularization type (penalty)**: L1 (Lasso) and L2 (Ridge)
- **Solver (solver)**: liblinear (suitable for small datasets)
- **Maximum iterations (max_iter)**: [100, 200, 500]

3.7 Tools and Libraries

Table 4: 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.
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.

Chapter 4: RESULT AND ANALYSIS

4.1 Model Evaluation Metrics

After training, the model's performance was evaluated on the test dataset. Metrics such as accuracy, precision, recall and f1-score were calculated from the confusion matrix to assess its effectiveness in classifying the news.

The confusion metric provides a detailed breakdown of the model's predictions, showing the number of true positive, true negative, false positive, and false negatives. The confusion matrix results provide a detailed view of the classification performance of the model. The confusion matrix is a table that summarizes how successful the classification model performs on labeled dataset. It provides the correct and incorrect classification for each class. One axis of the confusion matrix is predicted, and the other axis is the actual label.

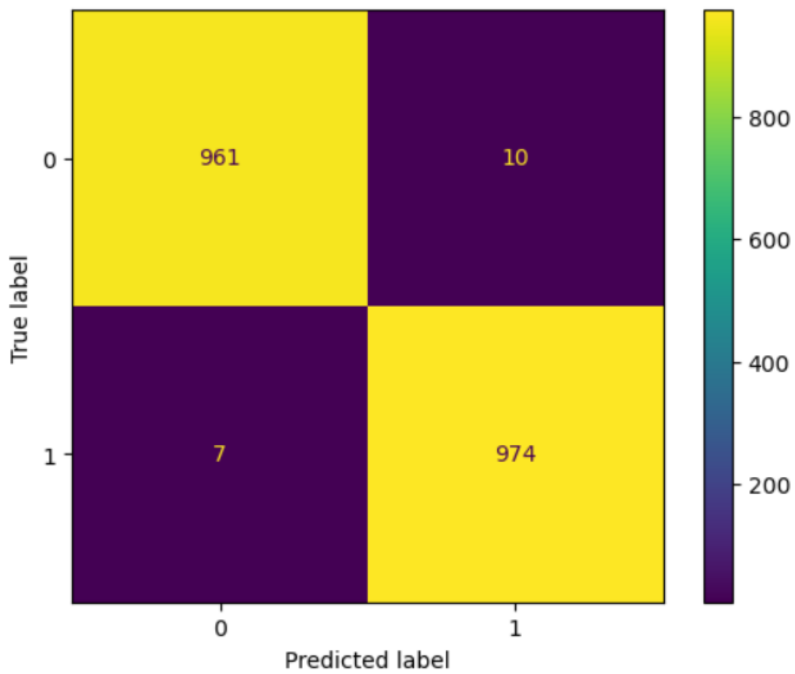

Figure 6: Confusion matrix of our model

Accuracy: Since we are working on a classification problem of classifying the Nepali News articles into different category. Accuracy will be a good evaluator for the models. This is most commonly use metric to evaluate how well the model predicts. Accuracy is the ratio of the number of correct predictions made against the total number of predictions made.

$$Accuracy = \frac{(True\ Positive + True\ Negative)}{Total\ Number\ of\ samples}$$

Precision: Precision calculates the ratio of correctly predicted positive instances to the total number of instances predicted as positive by the model. It assesses how well the model performs when it predicts a positive class.

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

Recall: Recall measures the model's ability to correctly identify positive instances (true positives) out of all the actual positive instances. It is also known as Sensitivity or True Positive Rate.

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

F1-Score: The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall, especially when you want to find a single metric to evaluate your model's performance.

$$\text{F1-Score} = \frac{(2*Precision*Recall)}{(Precision+Recall)}$$

ROC Curve:

The ROC curve is a visual representation of model performance across all thresholds. The ROC curve is drawn by calculating the true positive rate (TPR) and false positive rate (FPR) at every possible threshold (in practice, at selected intervals), then graphing TPR over FPR. A perfect model, which at some threshold has a TPR of 1.0 and a FPR of 0.0, can be represented by either a point at (0, 1) if all other thresholds are ignored, or by the following:

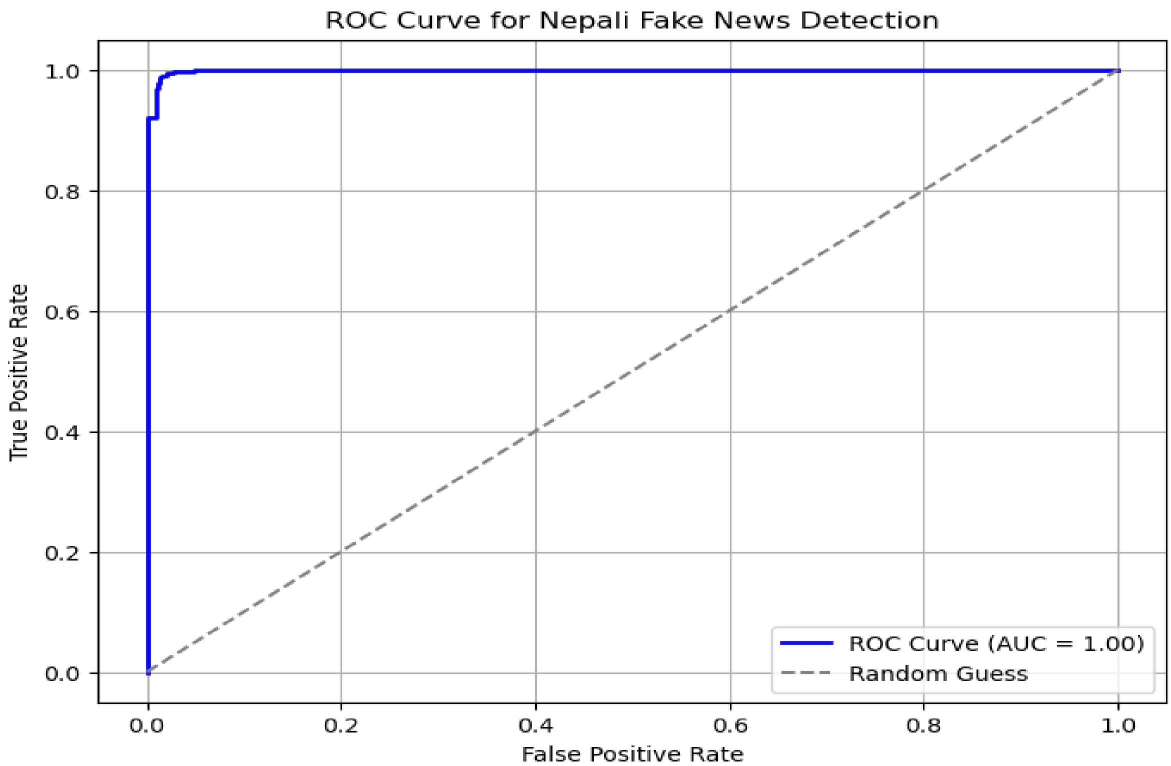

Figure 7: ROC curve of our model

4.2 Model Performance

The performance of the Nepali Fake News Detection model using Logistic Regression is measured based on key classification metrics. These metrics indicate that the model achieved a high level of accuracy in distinguishing between fake and real news.

- For Class 0 (Real News), the model demonstrated:

  Precision: 0.99, meaning that 99% of the news classified as real is actually real

  Recall: 0.90, indicating that the model correctly identified 99% of all real news in the dataset.

  F1-score: 0.99, showing a balance between precision and recall.

- For Class 1 (Fake News), the model achieved:

  Precision: 0.9, meaning that 98% of the news classified as fake is actually fake.

  Recall: 0.99, indicating that the model correctly identified 99% of all fake news in the dataset.

  F1-score: 0.99, demonstrating strong overall performance.

Overall Performance:

The macro average precision, recall, and F1-score are all 0.99, signifying balanced performance across both classes.

The weighted average precision is 0.90, while recall and F1-score remain at 0.99, indicating that the model maintains high performance even when considering class imbalances.

Chapter 5: CONCLUSION

The Nepali Fake News Detection system using Logistic Regression has been successfully implemented. This system is designed to classify news articles as either real or fake based on their content. The dataset consists of news articles collected from various Nepali news sources, and the project evaluates the effectiveness of machine learning techniques in detecting fake news.

Several preprocessing steps, such as stopword removal, symbol and number removal, and stemming, were applied to clean the text data. TF-IDF vectorization was used for feature extraction, and the processed features were then fed into the classification model for training. The proposed system analyzes the content of news articles to determine their authenticity.

Despite achieving high performance on the dataset, with precision, recall, and F1-scores close to 99%, the model still faces challenges in accurately identifying real news. This limitation arises due to unexplored linguistic complexities in the Devanagari script, such as variations in word usage, contextual meanings, and the lack of extensive labeled datasets. Addressing these challenges through improved linguistic preprocessing and a more diverse dataset could further enhance the model's accuracy in real-world scenarios.

Chapter 6: FUTURE ENHANCEMENT

- Incorporate automated data collection and preprocessing steps to keep our dataset updated.
- Study advanced techniques for Devanagari script tokenization, stemming, and lemmatization, as these can significantly improve text preprocessing for Nepali.
- Training our fake news detection model on a larger, more diverse dataset, to improve its accuracy.
- Develop a dynamic scraping mechanism to fetch real-time data and evaluate news articles against the model.

References

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

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

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

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

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