Tribhuvan University

Institute of Science and Technology

Seminar Report

On

NEPALI NEWS CLASSIFICATION USING NAIVE BAYES

Submitted to

Central Department of Computer Science and Information Technology

Tribhuvan University, Kirtipur

Kathmandu, Nepal

In the partial fulfilment of the requirement for Master's Degree in Computer Science and Information Technology (M. Sc. CSIT), Second Semester

Submitted by

Raju Shrestha

Roll No. 48/2079

May, 2024

Tribhuvan University

Institute of Science and Technology

SUPERVISOR'S RECOMMENDATION

This is to certify that Mr. Raju Shrestha has submitted the seminar report on the topic **"NEPALI NEWS CLASSIFICATION USING NAIVE BAYES"** for the partial fulfilment of Master's of Science in Computer Science and Information Technology, second semester. I hereby, declare that this seminar report has been approved.

…….

Supervisor

Asst. Prof. Arjun Singh Saud

Central Department of Computer Science and Information Technology

LETTER OF APPROVAL

This is to certify that the seminar report prepared by Mr. Raju Shrestha entitled **"NEPALI NEWS CLASSIFICATION USING NAIVE BAYES"** in partial fulfilment of the requirements for the degree of Master's of Science 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.

Evaluation Committee

……..……..	……..……
Asst. Prof. Sarbin Sayami	Asst. Prof. Arjun Singh Saud
(H.O.D)	(Supervisor)
Central Department of Computer Science and Information Technology	Central Department of Computer Science and Information Technology

……

(Internal)

ACKNOWLEDGEMENT

The success and final outcome of this report required a lot of guidance and assistance from many people and I am very fortunate to have got this all along the completion. I am very glad to express my deepest sense of gratitude and sincere thanks to my highly respected and esteemed supervisor **Asst. Prof. Arjun Singh Saud,** Central Department of Computer Science and Information Technology for his valuable supervision, guidance, encouragement, and support for completing this paper.

I am also thankful to **Asst. Prof. Sarbin Sayami**, HOD of Central Department of Computer Science and Information Technology for his constant support throughout the period. Furthermore, with immense pleasure, I submit by deepest gratitude to the Central Department of Computer Science and Information Technology, Tribhuvan University, and all the faculty members of CDCSIT for providing the platform to explore the knowledge of interest. At the end I would like to express my sincere thanks to all my friends and others who helped me directly or indirectly.

Raju Shrestha (48/2079)

ABSTRACT

With the exponential growth of digital content in Nepali language, the need for efficient classification techniques has become very important. This report proposes a Nepali news classification system that uses Naive Bayes algorithm. The primary objective is to categorize Nepali news articles into predefined categories such as politics, sports, entertainment, technology etc. This report discusses about the application of the Naïve Bayes algorithm for the classification of Nepali news articles into different categories collected from a major Nepali language newspaper published in Nepal. This project evaluates 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 articles. The proposed system first preprocesses the raw text data by tokenization, stop-word removal, and stemming to extract meaningful features. Subsequently, a Naive Bayes classifier is used to train on a labeled dataset comprising news articles from diverse categories. Evaluation of the system uses standard metrics such as accuracy, precision, recall, and F1-score are employed to assess the performance of the classification system. Results are analyzed to measure the effectiveness of the Naive Bayes algorithm in accurately classifying the Nepali news articles. This can facilitate Nepali-speaking users in accessing relevant news content efficiently, thereby enhancing their overall browsing experience. Additionally, the system can serve as a foundation for the development of more sophisticated natural language processing applications built for the Nepali language.

Keywords: *News Classification, Text Classification, Machine learning, Naive Bayes, Natural Language Processing*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI: Artificial Intelligence

Bi-LSTM: Bi-directional Long Short Term Memory

BOW: Bag of Words

CNN: Convolution Neural Network

CSV: Comma Separated Values

GRU: Gated Recurrent Unit

KNN: K-Nearest Neighbor

LSTM: Long Short Term Memory

MNN: Multilayer Neural Network

NB: Naïve Bayes

NLP: Natural Language Processing

NLG: Natural Language Generation

NLU: Natural Language Understanding

NLTK: Natural Language Toolkit

POS: Parts of Speech

RBF: Radial Basis Function

RNN: Recurrent Neural Network

SVM: Support Vector Machine

TF-IDF: Term Frequency Inverse Document Frequency

CHAPTER 1: INTRODUCTION

1.1 Overview

In the present days, Internet has produces vast amount of textual data, which can be termed in other words as unstructured data. Internet and corporate spread across the globe produces textual data in exponential growth, which needs to be shared, on need basis by individuals. If the data generated is properly organized, classified then retrieving the needed data can be made easily with least efforts. Hence the need of automatic methods to organize and classify the news articles into different category is very important due to such exponential growth of information. Automatic classification refers to assigning the news articles to a set of pre-defined category such as education, business, sport, health and technology etc. based on the textual content.

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 articles 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 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.

News classification for the Nepali language is a challenging problem due to complexity of the language. Limited research have been carried out in this domain. This work proposed a supervised Machine Learning based framework for classifying Nepali news articles into different category.

1.3 Objective

The main objectives of this seminar report is to classify the Nepali news articles into pre-defined categories using Naive Bayes algorithm.

CHAPTER 2: BACKGROUND STUDY AND LITERATURE REVIEW

2.1 Background Study

Natural Language Processing (NLP) is a dynamic field of artificial intelligence (AI) that focuses on enabling computers to understand, interpret, and generate human language. . Human language consists of words and sentences in natural form and NLP is used to extract the information. NLP processes or analyzes written or spoken language. Natural Language Processing is the analysis of linguistic data in the form of text as documents or sentences using computational methods. It helps to represent the unstructured text into structured text data using computational linguistics. NLP research try to collect information on how human understand and use different natural languages that helps machine understand and manipulate natural languages.

Natural language processing encompasses a wide range of tasks, from basic text processing like language translation, text summarization, morphological analysis and sentiment analysis to more complex applications such as chatbots, speech recognition, and language generation. NLP algorithms employ techniques from machine learning, linguistics, and computer science to bridge the gap between human communication and computational systems. With its rapid advancements and real-world applications across industries like healthcare, finance, and customer service, NLP plays a vital role in re-shaping how humans interact with machines and access information, making it a vital area of study and innovation in today's digital age.

Natural language processing is the field of machine learning which is concerned with training machines to understand and generate results like humans is done by using natural languages. It has two components Natural Language Understanding (NLU) and Natural Language Generation (NLG).

Natural language understanding (NLU) is deals with machine reading comprehension. Natural language understanding is considered an AI-hard problem. The system needs to disambiguate the input sentence to produce the machine representation language. NLU is used in the automated reasoning, machine translation, question answering, news gathering, text categorization, voice activation, archiving, and large scale content analysis. Natural language generation (NLG) is a software process that produces natural language as output. NLG is

complementary to natural-language understanding. The system needs to make decisions about how to put a representation into words.

The Natural Language Processing in Nepali language has not a long history as NLP in Nepali text become started from early 2000s AD. But the works are in high number now days. The Nepali NLP works are broadly done on three categories namely Rule based, Machine Learning based and deep learning based.

2.1.1 News Classification

Classification is the process of predicting the class of data based on label. Labeled data has previously categorized into one or more class. Unlabeled data is data that has not yet been labeled. It is a process of automatically assigning a label to an unlabeled data. 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 20 different types of news has been classified like business, sports, entertainment, political, literature and many more.

2.1.2 Naive Bayes Classification

Bayes Theorem describes the probability of an event which is based on prior knowledge of conditions that might be related to the event. It is named after Thomas Bayes. Mathematically Bayes theorem is stated as,

$$P(A|B) = \frac{\mathrm{P(B|A)} * \mathrm{P(A)}}{\mathrm{P(B)}}$$

where,

P (A) = Probability of A

P (B) = Probability of B

P (A|B) = Probability of A given B

P (B|A) = Probability of B given A

Naïve Bayes algorithm is comprised of two words “Naive”, an independent occurrence of features and “Bayes”, an algorithm that depends on Bayes theorem. It is one of the supervised machine learning algorithms. It is a probabilistic classifier that calculates a set of probabilities by counting the frequency. Naive Bayes is based on Bayes Theorem, which help to compute conditional probabilities of the occurrence of two events, based on the probabilities of the occurrence of each individual event. A Naive Bayes classifier makes the assumption that all the input features are independent of each other. It is a fast, easy, computationally less time taken method of classification. It is also used in spam filtering, sentiment analysis, recommendation system, multiclass classification etc.

Multinomial Naive Bayes is a very popular and efficient machine learning algorithm. It is commonly used for text classification tasks where we need to deal with discrete data like word counts or term frequencies in documents. Multinomial model is used for multinomial distributed data. Features having a given term appear number of times, multinomial Naïve Bayes Model is used. It is mainly used in document classification problems, where a particular document is categorize to different category such as Sports, Politics, education, etc.

2.2 Literature Review

Text classification have become the growing area in the field of natural language processing. Supervised machine learning algorithm like Naïve Bayes algorithm plays vital role in the text classification. There are many research carried out for text classification task.

In paper [1] discuss about different text classification techniques, including bags-of-words, bag-of-n-grams, and convolutional neural networks (CNN), for the task of classifying newspaper articles. Here, CNN achieves the highest accuracy on compared to the traditional models such as SVM, Naïve Bayes and KNN.

In paper [2] presents an approach for classifying news articles using machine learning techniques, specifically focusing on the application of Naive Bayes classifier and TF-IDF

vectorization. This paper presents the results of Naive Bayes classification with and without TF-IDF vectorization. The findings indicate that Naive Bayes with count vectorization achieves higher accuracy compared to Naive Bayes with TF-IDF vectorization.

In paper [3] presents an approach for classifying news articles using machine learning techniques, specifically focusing on the application of Naive Bayes classifier and TF-IDF vectorization. The paper presents the results of Naive Bayes classification with and without TF-IDF vectorization. The findings indicate that Naive Bayes with count vectorization achieves higher accuracy compared to Naive Bayes with TF-IDF vectorization.

In paper [4] focuses on the task of Nepali news classification using different deep learning algorithms, including Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), and Transformer models. The authors collected a dataset of around 200,000 news articles from various Nepali news portals, covering 17 different news categories. They preprocessed the data, including text cleaning, tokenization, stop word removal, and subword tokenization. The models were trained and evaluated using an 80-20 train-test split, with 20% of the training data used for validation. The results show that the Transformer model outperformed the LSTM and Bi-LSTM models, achieving a training F1-score of 96.54% and a testing F1-score of 95.45%. The Bi-LSTM model achieved a training F1-score of 94.58% and a testing F1-score of 93.65%, slightly better than the unidirectional LSTM model.

In paper [5] explores the use of deep learning techniques, specifically Recurrent Neural Network (RNN) and Multilayer Neural Network (MNN), for the classification of Nepali text documents. The results show that the RNN model outperformed the MNN model, with the highest accuracy of 63% for RNN compared to 48% for MNN. The RNN model consistently achieved higher performance across the evaluation metrics.

In paper [6] discusses the classification of Nepali news articles into different categories using deep learning techniques. The increasing popularity of online news portals and the large volume of news articles produced regularly necessitate an automated system to organize and categorize the content. The LSTM model achieved the highest accuracy of 95.36%, outperforming the CNN (93.97%) and DNN (90.75%) models. The LSTM model also showed better precision, recall, and F1-score compared to the other two models.

In paper [7] focuses on the problem of automated news classification for the Nepali language. It examines the use of various machine learning techniques, including Naive Bayes, Support Vector Machines (SVMs), and Neural Networks, for classifying Nepali news articles into

predefined categories. The results show that the SVM with RBF kernel achieved the highest average accuracy of 74.65%, followed by linear SVM with 74.62%, Multilayer Perceptron Neural Networks with 72.99%, and Naive Bayes with 68.31%. The SVM-based models also outperformed the other techniques in terms of precision, recall, and F-score.

In paper [8] focuses on multi-class text classification for the Nepali language using traditional machine learning algorithms and the latest deep learning algorithms. The results show that the traditional machine learning algorithms, such as Perceptron, achieved the highest accuracy of 78.561% on the testing dataset. Among the deep learning algorithms, GRU network achieved an accuracy of 77.44%, which was lower than the traditional machine learning models.

In paper [9] focuses on the comparison of different classification algorithms for classifying Nepali news articles. They compare the performance of four classification algorithms (SVM-RBF Kernel, SVM-Poly Kernel, NB Multinomial, and Random Forest) for the task of Nepali news classification. The results show that the SVM-Poly Kernel algorithm outperformed the other three algorithms, achieving an accuracy of 82.76%, precision of 82.9%, recall of 82.8%, and F-measure of 82.7%. The SVM-RBF Kernel, NB Multinomial, and Random Forest algorithms had lower performance compared to the SVM-Poly Kernel.

CHAPTER 3: METHODOLOGY

Nepali News Classification uses supervised machine learning algorithm i.e. Naïve Bayes. This requires labeled data for training the classifier. The detail architecture of the system is presented below.

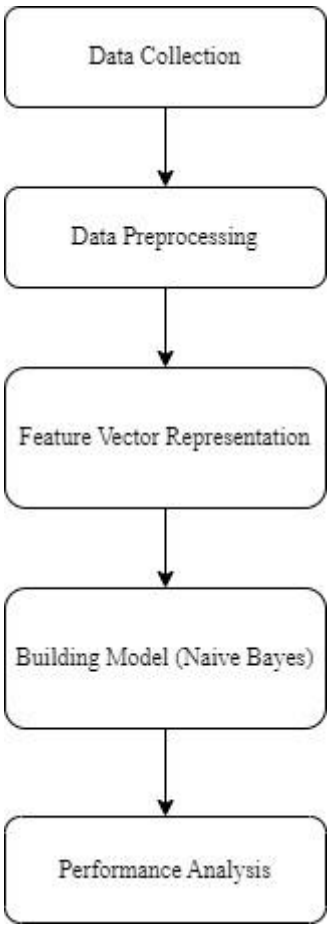

Figure 3.1: System Architecture

3.1 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 along with different modifiers and half forms. The Nepali character set present in Nepali language is given in Table 1.

Table 1: Nepali Character Set

Numerals	०, १, २, ३, ४, ५, ६, ७, ८, ९
Consonants	क, ख, ग, घ, ङ, च, छ, ज, झ, ञ, ट, ठ, ड, ढ, ण, त, थ, द, ध, न, प, फ, ब, भ, म, य, र, ल, व, श, ष, स, ह, क्ष, त्र, ज्ञ
Vowels	अ, आ, इ, ई, उ, ऊ, ए, ऐ, ओ, औ, अं, अः

The dataset used for Nepali News Classification Using Naïve Bayes is Nepali News dataset 20 categories which is collected from the publicly available source i.e. from Kaggle, an online repository known for its diverse collection of data. The dataset consists of 4964 data in csv format. These dataset consists of news articles collected from various online news portal such as setopati.com, onlinekhabar.com, ekantipur.com etc. The dataset consists of training data and is labeled into different categories. The sample dataset used in this project is shown below.

	text	category
0	काठमाडौंमा पहिलो पटक स्ट्रबेरीको व्यवसायिक खे...	Agriculture
1	जिल्लाका किसानले लगाएको अदुवामा गानो कुहिने, ग...	Agriculture
2	काभ्रेपलाञ्चोकमा कृषकले एसआरआई प्रविधिमा गरेको...	Agriculture
3	राजधानीमा यतिबेला तरकारीको मूल्य आकासिएको छ। क...	Agriculture
4	पाल पोल्ट्री तथा लाइभस्टक क्षेत्रको समग्र विका...	Agriculture
5	माग अनुसारको आपूर्ति नभएपछि तरकारीको मुल्य दोब...	Agriculture
6	नेपाल र चीनबीचको सम्झौतासँगै शुरु भएको परियोजन...	Agriculture
7	यस वर्ष अम्बामा देशैभर अम्बामा रोग देखा पऱ्यो...	Agriculture
8	लैँचीको कारोबार ठप्प भएपछि चाडबाडका लागि खर्च ...	Agriculture
9	बदलिँदो मौसमका कारण मोरङमा उन्नत जातको धानबाली...	Agriculture

Figure 3.2: Nepali News Dataset

The News dataset consists of category of news article and the number of news articles per category. The distribution of news articles in different category is shown in Table 2.

Table 2: News Category & Number of Documents

S.N	Category	Number of Documents
1	Agriculture	100
2	Automobiles	95
3	Bank	417
4	Blog	209
5	Business	142
6	Economy	500
7	Education	85
8	Employment	154
9	Entertainment	500
10	Health	31
11	Interview	229
12	Literature	102
13	Migration	111
14	Opinion	500
15	Politics	500
16	Society	253
17	Sports	500
18	Technology	110
19	Tourism	214
20	World	212
Total		4964

3.2 Data Preprocessing

Data preprocessing is important step in NLP. It cleans the text data and makes it ready for machine learning model. Preprocessed data helps to improve the performance of classifier and speedup the classification process. The preprocessing techniques include stop words removal, symbol and number removal and stemming.

1. Stop words removal: Stop word 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 lists for English and other language are easily available but there is not any standard stop words list for the Nepali language. Some of the stop words are given as follows.

Determiners: यो, त्यो, ती, हरेक, प्रत्येक, सबै, केही, को, कुन, कति, जो, जसरी, जुन

Conjunctions: र, तथा, तर, किन्तु, परन्तु, कनकि, पनि

Postpositions: हरु, ले, लाई, बाट, द्वारा, वारि, पारि

These stop words need to be remove 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.

2. 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

3. Word Stemming: 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.

3.3 Feature Vector Representation

Feature vector representation is representation of text into vector space. It converts text into numerical representation. It helps to improve the scalability, efficiency and accuracy of model.

Some of the popular and simple methods to extract features from text are Bag of words/Count Vector, TF-IDF, Word2vec, Glove etc.

3.3.1 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is widely used feature vector representation technique for the text analysis in natural language processing. It is a statistical method to find importance of word 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.

1. **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;

$$\textit{Term Frequency (TF)} = \frac{Number\ of\ times\ \ term\ occurrences\ in\ a\ document}{Total\ Number\ of\ words\ in\ a\ document}$$

2. **Inverse Document Frequency (IDF):** It is the number of documents that contain a term in the collection of document. It is a document-level statistic that gives 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.

$$\textit{Inverse Document Frequency (IDF)} = \log_e\{\frac{Total\ number\ of\ documents}{Number\ of\ documents\ which\ are\ having\ term}\}$$

Formula to calculate complete TF-IDF value is,

$$\textit{TF-IDF} = \textit{TF} * \textit{IDF}$$

3.4 Naive 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.

CHAPTER 4: IMPLEMENTATION AND TESTING

This section describes the tools and technologies used for classifying the News articles into different category using Naïve Bayes Algorithm. Nepali News Classification Using Naïve Bayes algorithm model is implemented in Python programming language and Jupyter notebook is used as software environment. Python is a high-level, general purpose programming language. It has a great support for libraries for implementing machine learning algorithm. The libraries used for this work are as follows:

- **Pandas:** Pandas is an open source Python library that provides fast, powerful, flexible, high performance, easy to use data structures and data analysis tools. It also allows various data manipulation operations.
- **NumPy:** NumPy is a general purpose array processing Python library used for array variable and numerical computing.
- **Matplotlib:** Matplotlib is a plotting library for 2D graphics in python programming language.
- **Scikit learn:** Scikit-learn is the simple and efficient tools for predictive data analysis. It provides the tools for classification, regression, clustering, preprocessing and model selection.
- **Natural Language Toolkit (NLTK):** NLTK is a toolkit used for NLP in Python. It can be used to performing different NLP task such as tokenization, stop words removal, stemming, lemmatization, parse tree generation, parts of speech (POS) tagging etc.

4.1 Description of Major Function

The major function in the application are:

4.1.1 Preprocessing

The data was preprocessed with different NLP technique such as stop words removal, removing numbers and punctuation, word stemming etc.

```
# Text Preprocessing
from snowballstemmer import NepaliStemmer

def preprocess_text(cat_data, stop_words, punctuation_words):
    stemmer = NepaliStemmer() # initialize the Nepali stemmer
    new_cat = []
    noise = "1,2,3,4,5,6,7,8,9,0,०,१,२,३,४,५,६,७,८,९".split(",")
    for row in cat_data:
        words = row.strip().split(" ")
        nwords = ""
        for word in words:
            # apply Nepali stemming to the word
            if word not in punctuation_words and word not in stop_words:
                word = stemmer.stemWord(word)
                is_noise = False
                for n in noise:
                    if n in word:
                        is_noise = True
                        break
                if not is_noise and len(word) > 1:
                    word = word.replace("(","")
                    word = word.replace(")","")
                    nwords += word + " "

        new_cat.append(nwords.strip())
    return new_cat

title_clean = preprocess_text(["\nसगरमाथा चुचुरोमा पुग्ने ९ शेर्पा टोलीको नेतृत्व लेख्नु चाहनुहुन्छ गरेको छ ।"], stop_words, punctuation_words)
print(title_clean)
```

Figure 4.1: Data Preprocessing

4.1.2 Feature Extraction

Feature extraction plays major role in the classification system and it is heart of the classification system. A good feature sets should represent characteristic of a class that helps to distinguish it from other classes, while remaining invariant to characteristic differences within the class. Hence, to improve the accuracy of the classifier, it is necessary to identify a set of "good" features for object representation. To create improved features, measurement of various object properties are carried out to identify good features from a set of raw features.

The data was then performed feature vector representation by using TF-IDF method. It gives the weightage of words and reflect how important a word is in a document. The Nepali text words were used as features for data analysis. The vector representation of text was then used for model training.

```
# Convert text the data into numerical features using CountVectorizer
from sklearn.feature_extraction.text import CountVectorizer

vectorizer = CountVectorizer(ngram_range=(1, 2)).fit(X_train)
X_train_vectorized = vectorizer.transform(X_train)
print(X_train_vectorized)
X_train_vectorized.shape
```

```
# Transform data using TFIDF transformer
from sklearn.feature_extraction.text import TfidfTransformer

tfidf_transformer=TfidfTransformer()
X_train_tfidf=tfidf_transformer.fit_transform(X_train_vectorized)
X_train_tfidf.shape
```

Figure 4.2: Feature Extraction

4.1.3 Model Training

In Naïve Bayes method, multinomial Naïve Bayes was used for training the model. This is one of the classic Naïve Bayes algorithm that implements the multinomially distributed data. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

```
# Train the Naive Bayes Classifier
from sklearn.naive_bayes import MultinomialNB

model = MultinomialNB(alpha=0.1)
model.fit(X_train_vectorized, Y_train)
model.score(X_train_vectorized, Y_train)
```

Figure 4.3: Model Training

4.2 Testing

Among the total data 80% of the data is used for training and 20% is used for testing. For testing hit and trial method is followed and the testing module from Scikit-learn is used for testing.

```
# Split the data into training and testing sets
from sklearn.model_selection import train_test_split

X_train, X_test, Y_train, Y_test = train_test_split(data["text"], data["target"], test_size=0.2, random_state=1)
```

Figure 4.4: Train-Test Split

4.3 Model Evaluation

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.

```
# Calculate weighted accuracy
correct_predictions = np.sum(np.diag(conf_matrix))
total_instances = np.sum(conf_matrix)
weighted_accuracy = correct_predictions / total_instances

# Print the results
print("Weighted Accuracy :", weighted_accuracy*100, '%')
```

```
# Calculate macro/micro precision, recall & f1-score
from sklearn.metrics import precision_score, recall_score, f1_score

precision_macro = precision_score(Y_test, y_predicted, average='macro')
precision_micro = precision_score(Y_test, y_predicted, average='micro')

recall_macro = recall_score(Y_test, y_predicted, average='macro')
recall_micro = recall_score(Y_test, y_predicted, average='micro')

f1_macro = f1_score(Y_test, y_predicted, average='macro')
f1_micro = f1_score(Y_test, y_predicted, average='micro')

# Print the results
print("Macro Precision Score :",  precision_macro*100, '%')
print("Micro Precision Score :", precision_micro*100, '%')

print("Macro Recall Score :", recall_macro*100, '%')
print("Micro Recall Score :", recall_micro*100, '%')

print("Macro F1 Score :", f1_macro*100, '%')
print("Micro F1 Score :", f1_micro*100, '%')
```

Figure 4.5: Model Evaluation

CHAPTER 5: RESULT AND FINDINGS

5.1 Evaluation Metrics

Confusion Metrics: The confusion metrics 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.

```
11   0   0   0   1   0   0   0   0   0   0   1   0
 0  14   0   3   0   0   0   0   1   0   0   3   0
 0   2  73   1   1   0   0   0   0   0   1   0   2
 0   0   2  19   0   0   0   0   0   0   0   0   0
 2   2   5   2   5   0   0   0   1   0   0   2   0
 0   0   2   2   1  12   0   0   2   0   0   0   1
 1   0   1   2   1   0  12   0   1   0   0   2   1
 0   0   0   0   0   0   0  18   4   0   0   0   1
 0   0   1   1   0   0   0   0  36   0   0   0   0
 0   1   2   1   0   1   0   0   1  55   0   0   2
 0   0   1   1   0   0   0   0   0   0  53   0   0
 1   0   0   0   2   1   0   0   0   0   0  28   0
 1   0   1   3   0   1   0   0   0   0   0   0  15
```

Figure 5.1: Confusion Matrix

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 prediction 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.

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

5.2 Model Performance

The performance of Nepali News Classification Using Naïve Bayes is measured. These metrics indicate that the model achieved a quit low accuracy in correctly classifying News into different category. The precision score reflects the model's ability to correctly classify the news. The recall score indicates its ability to classify all the actual positive news present in the dataset. F1 score indicates the model's ability to achieve both accurate positive news classification and comprehensive capturing of actual positive news.

- Accuracy (Weighted): 82.20 %
- Precision (Macro): 79.03 %
- Precision (Micro): 82.20 %
- Recall (Macro): 76.31 %
- Recall (Micro): 82.20 %
- F1-Score (Macro): 76.25 %
- F1-Score (Micro): 82.20 %

CHAPTER 6: CONCLUSION

Nepali News Classification Using Naïve Bayes algorithm is successfully implemented in Python programming language. A system has been built for classifying the content of the news article into different categories collected from a major Nepali language newspaper published in Nepal. This project evaluates some widely used machine learning techniques i.e. Naïve Bayes for automatic Nepali News classification. After collecting data various preprocessing steps such as stop words removal, symbol and number removal and stemming were performed. TF-ID vectorization is used for feature extraction and then the extracted features are passed to the classification system for training the model. The purposed system classifies the news by analyzing the content of the news article. Finally the weighted average accuracy of the model obtained on test data is 82.20 %.

REFERENCES

[1] J. Sreedevi, M. Rama Bai and C. Reddy, "Newspaper Article Classification using Machine Learning Techniques," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 5, March 2020.

[2] S. D. Mahajan and D. D. Ingle, "News Classification Using Machine Learning," *International Journal of Innovative Science and Research Technology*, vol. 6, no. 5, May – 2021.

[3] M. Sundarababu, M. Ch.Chandra, M. Suthar, C. Harsha, L. Juveria and B. Blessy, "NEWS CLASSIFICATION USING MACHINE LEARNING," *JETIR*, vol. 7, no. 3, 2020.

[4] S. S. Wagle and . S. Thapa, "Comparative Analysis of Nepali News Classification using LSTM, Bi-LSTM and Transformer Model," *Proceedings of 10th IOE Graduate Conference*, vol. 10, October 2021.

[5] S. Subba, N. Paudel and T. B. Shahi, "NEPALI TEXT DOCUMENT CLASSIFICATION USING DEEP NEURAL NETWORK," *Tribhuvan University Journal TRIBHUVAN*, vol. 33, June 2019.

[6] S. K. Thapa and S. Pokhrel, "Nepali News Document Classification using Global Vectors and Long Short Term Memory," *Proceedings of 9th IOE Graduate Conference*, vol. 9, March 2021.

[7] T. B. Shahi and A. K. Pant, "Nepali News Classification using Naïve Bayes, Support Vector Machines and Neural Networks," *International Conference on Communication, Information & Computing Technology (ICCICT),* feb 2018.

[8] O. M. Singh, "Nepali Multi-Class Text Classification," *Department of CSEE University of Maryland, Baltimore County*, December 20 2018.

[9] K. Acharya and S. Shakya, "An Analysis of Classification Algorithms for Nepali News," *International Journal of Innovative Science, Engineering & Technology*, vol. 7, no. 7, July 2020.