Tribhuvan University
Institute of Science and Technology

PERFORMANCE ANALYSIS OF NAÏVE BAYES AND SUPPORT VECTOR MACHINE ALGORITHM ON CLASSIFICATION OF NEPALI OPINION TEXT

Dissertation

Submitted to
Central Department of Computer Science and Information Technology
Tribhuvan University, Kirtipur
Kathmandu, Nepal

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

Submitted by
Nishchhal Shrestha
May, 2022

Tribhuvan University
Institute of Science and Technology

PERFORMANCE ANALYSIS OF NAÏVE BAYES AND SUPPORT VECTOR MACHINE ALGORITHM ON CLASSIFICATION OF NEPALI OPINION TEXT

Dissertation

Submitted to

Central Department of Computer Science and Information Technology
Tribhuvan University, Kirtipur
Kathmandu, Nepal

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

Submitted by

Nishchhal Shrestha

May, 2022

Supervisor

Asst. Prof. Bikash Balami

Tribhuvan University
Institute of Science and Technology
Central Department of Computer Science and Information Technology

Student's Declaration

I hereby, declare that I am the only author of this work and that no sources other than the listed here have been used in this work.

…………………………………..
Nishchhal Shrestha

Date: ……………………..

Tribhuvan University
Institute of Science and Technology
Central Department of Computer Science and Information Technology

SUPERVISOR'S RECOMMENDATION

I hereby recommend that this dissertation prepared under my supervision by **Mr. Nishchhal Shrestha** entitled **"PERFORMANCE ANALYSIS OF NAÏVE BAYES AND SUPPORT VECTOR MACHINE ALGORITHM ON CLASSIFICATION OF NEPALI OPINION TEXT"** in partial fulfillment of the requirement for the degree of Master's of Science in Computer Science and Information Technology be processed for the evaluation.

……………………………………

Asst. Prof. Bikash Balami
Central Department of Computer Science
and Information Technology (CDCSIT)
Tribhuvan University

Tribhuvan University
Institute of Science and Technology
Central Department of Computer Science and Information Technology

LETTER OF APPROVAL

We certify that we have read this dissertation and in our opinion it is satisfactory in the scope and quality as a dissertation in the partial fulfillment for the requirement of Master's Degree in Computer Science and Information Technology.

Evaluation Committee

…………………………………

Asst. Prof. Sarbin Sayami
Central Department of Computer Science and Information Technology (CDCSIT)
Tribhuvan University
(Head of Department)

…………………………………

Asst. Prof. Bikash Balami
Central Department of Computer Science and Information Technology (CDCSIT)
Tribhuvan University
(Supervisor)

…………………………………

(External Examiner)

…………………………………

(Internal Examiner)

Date: …………………………………

ACKNOWLEDGEMENT

First of all, I would like to express my sincere and hearty gratitude to my supervisor Asst. Prof. Bikash Balami for the guidance, frequent monitoring, continuous inspiration, co-operation, valuable suggestions and constant support during the thesis work.

I am greatly thankful to Asst. Prof. Sarbin Sayami, Head of Central Department of Computer Science & Information Technology (CDCSIT), Kirtipur for providing valuable suggestion, support and inspiration to complete this dissertation.

I am thankful to Prof. Dr. Subarna Shakya, Assoc. Prof. Nawaraj Paudel, Asst. Prof. Dhiraj Kedar Pandey, Asst. Prof. Lalita Sthapit, Asst. Prof. Jagdish Bhatta, Asst. Prof. Tej Bahadur Shahi, Asst. Prof. Bal Krishna Subedi and Asst. Prof. Ram Krishna Dahal of Central Department of CSIT TU for their kind suggestion and inspiration during the course of this work.

I am also thankful to Mamata Chalise and Roshan Pulami staff of CDCSIT, Tulasi Kumari Sapkota (Librarian of Central Department of CSIT), Central Library Kirtipur and Nirmal Home Library Tansen Palpa.

It would not have been possible to complete this dissertation without the valuable suggestion, guidance, support and inspiration from various institutions, individuals and friends.

I am grateful to Ms. Kesh Laxmi Shrestha, my parents Mr. Nirmal Shrestha and Mrs. Renuka Shrestha for their support and encouragement. I would like to express my gratitude to my brother and sisters for continuous support during my thesis work.

I would like to thank Mr. Shyam Prasad Shrestha for computer setting. At last, my acknowledgement goes to all those who more or less contributed me in completing this work.

Nishchhal Shrestha
May, 2022

ABSTRACT

Opinion is a subjective expression of individual on something. These are views, emotions or sentiments. The opinion helps individual and organization to make decision about the certain things. The opinion classification is the process of analyzing the view or opinion using the natural language processing techniques. The Naïve Bayes and Support Vector Machine (SVM) algorithm are supervised machine learning algorithm for classification. Most of the researches in opinion classification are done in English language but it is important to perform the opinion classification in Nepali language as the amount of data in Nepali is increasing rapidly in the form of blog, review, opinion column in newspaper. Nepali sentences were collected from the opinion section of different online portal of national newspaper in this study. The python programming language was used for implementing both algorithms with NLTK library and output were analyzed on the basis of performance metrics. The accuracy of SVM is 85% which is higher than accuracy of Naïve Bayes algorithm i.e. 83% on preprocessed the data. The accuracy of both algorithms was improved after preprocessing as compared to without preprocessing the data. The Study concluded SVM model was the best model with higher values of performance metrics and is recommended for opinion classification of Nepali text data over the Naïve Bayes algorithm.

Keyword: *Opinion Classification, Natural Language Processing, Support Vector Machine, Naïve Bayes Classifier, Nepali Opinion Text Dataset*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CSV	:	Comma Separated Values
FN	:	False Negative
FP	:	False Positive
ML	:	Machine Learning
NB	:	Naive Bayes
NLG	:	Natural Language Generation
NLP	:	Natural Language Processing
NLTK	:	Natural Language Tool Kit
NLU	:	Natural Language Understanding
SVM	:	Support Vector Machine
TF-IDF	:	Term Frequency-Inverse Document Frequency
TN	:	True Negative
TP	:	True Positive

CHAPTER 1
INTRODUCTION

1.1. Introduction

An ability of computer to understand the language as human in written or speech format is known as Natural language processing. The classification is the process of categorizing the data into two or more classes by constructing the model. It is the supervised learning algorithm which has labeled data. The model is performed on unlabeled data that classify into the predefined classes. The classifications of text in natural language have the significant importance in the development of language understanding by machine. Opinion is a view or judgments formed about something which may or may not based on fact or knowledge. It is individual's judgment or perspective about something. Opinion is a belief, judgment or way of thinking about particular things [1]. Opinion is the subjective expression which describes people's views, emotions and sentiments towards entities and their properties.

Opinion classification is the process of classifying the opinion in written text or review of users. It is useful to define the user view as opinion or not. Analysis of opinion for particular incident, politics, movies, sports, current issues, product, news are beneficial to government, companies, institutions and individuals for making strategy, marketing & advertising.

The supervised classification algorithm are of different types such as support vector machine, artificial neural network, logistic regression, naïve bayes, k-nearest neighbor, decision tree. Naïve Bayes and Support Vector Machine algorithm are most used for text classification for natural language. The Naive Bayes algorithm is a supervised machine learning algorithm used for categorization problem. Thomas Bayes (1702-1761) founded the Bayes theorem and Naive Bayes Classifier is based on that theorem. The Bayes theorem is a mathematical formula for determining conditional probability. Conditional probability is the probability of an event or outcome occurring, based on a previous event or outcome occurring. Naive Bayes classifier is a classification algorithm used for binary and multi-class problems. This technique is easy to understand for binary and categorical values. It is called a Naïve Bayes classifier because Bayesian classifier makes a simplifying (naive) assumption about how the features interact. The Support Vector Machine (SVM) is one of the supervised machine learning algorithms used for different types of problem like classification, regression problem. The Support Vector Machine principle was first introduced by Vapnik et.al. on the basis of statistical learning theory [2]. SVM algorithm works with finding a hyper plane that divide the classes. SVM uses the concept of Support Vectors that are the closest points from the hyper plane. SVM is basically binary classifier but it can be extended into multiclass classifier by using different heuristic methods such as one-vs-one and one-vs-rest. Kernel function like linear, RBF, polynomial, sigmoid etc. are used for decision making in the SVMs.

In Nepal, the opinion giving tradition is increasing in recent time. These opinions are in written text and speech. Nepali opinion text which is increasing in online and offline media is also high. These are on social media platform, blogging site and opinion column in newspaper.

1.2. Problem Statement

The increase in the amount opinion in Nepali text without classifying (either opinion or not opinion) is meaningless. These are high in volume and unstructured in nature. So these opinion data are needed to be preprocessed, classify and make them to meaningful. So opinion classification play important role. Opinion classification for the Nepali language is a challenging problem due to complexity of the language. Limited works have been done so far in this domain. This work proposed a supervised Machine Learning based framework for classifying opinions in Nepali text.

This research work classified collected Nepali Text data into one of the classes i.e. opinion or not opinion which helps in future in the domain of opinion in Nepali text. This also added significant advantages in research of opinion in Nepali language.

1.3. Objectives

The main objectives are;

- To study the opinion classification on Nepali text using Naïve Bayes and Support Vector Machine Algorithm.
- To compare results and evaluate the performance of Naïve Bayes and Support Vector machine algorithm using accuracy, precision, recall.

1.4 Dissertation Organization

This dissertation is outlined on following manner.

Chapter 1 consists of introduction, problem statement and objectives.

Chapter 2 describes about the background study for the research and literature review of the related work by different authors.

Chapter 3 includes the overview of the methodology and implementation of Naïve Bayes and Support Vector Machine algorithm for this research work.

Chapter 4 contains the result and performance analysis of Naïve Bayes and Support Vector Machine algorithm on Nepali opinion data.

Chapter 5 concludes with the summarizing the finding and future recommendations.

CHAPTER 2

BACKGROUND STUDY AND LITERATURE REVIEW

2.1 Background Study

2.1.1 Machine Learning

Machine Learning is a part of Artificial Intelligence that is used to make computer learn. According to Arthur Samuel, ‘Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed’. Machine learning (ML) algorithms train the data to predict the unknown data [3]. Machine Learning can be applied in different areas such as automatically classifying news articles, detecting tumors in brain scans, summarizing documents automatically, creating a chatbot (personal assistant), forecasting, client segmentation etc. The algorithms used for machine learning are supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.

2.1.1 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a subfield of artificial intelligence that deals with 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 [4]. NLP research try to collect information on how human understand and use different natural languages that helps machine understand and manipulate natural languages [5].

Supervised classifiers can be used different NLP tasks as document classification, part-of-speech tagging, sentence segmentation, dialogue act type identification, determining entailment relations, text summarization, sentiment analysis, machine translation, text and speech processing, morphological analysis [6]. The process of training machines to understand and generate results like humans is done by using natural languages processing. It has two components Natural Language Understanding (NLU) and Natural Language Generation (NLG) [7].

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.

NLP has five phases namely lexical analysis, syntactic analysis, semantic analysis, discourse integration pragmatic analysis [8]. Lexical Analysis involves identifying and analyzing the structure of words. It divides the text into paragraph, sentences and words. Syntactic Analysis is also known as parsing. It analyzes the words in grammatical manner that helps to find the relationship between words. The sentences are taken as an input and give a grammatical representation as an output. Semantic Analysis contain finding of meaning from the text. It converts syntactic representation into a meaningful representation. Semantic analysis is done by word sense determination and sentence level analysis. Discourse Integration involves the analysis of preceding sentences that helps in finding the meaning of current sentence. It also helps in finding the meaning of immediately succeeding sentence. Pragmatic Analysis involves the derivation of aspects of real world knowledge. Pragmatics includes the aspects of meaning that depends on the context or fact. These aspects include pronoun and referring expression, logical inferences, discourse structure, handling pronouns, handling ambiguity.

Application of NLP

Natural Language Processing plays significant role in every field due to its broader application area. Some of application areas of NLP are;

- Machine Translation
- Text categorization
- Extracting data from text
- Information retrieval
- Question Answering
- Sentiment Analysis
- Text to speech
- Speech recognition

Natural Language Processing in Nepali Language

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 different types of NLP research on Nepali language is done basically three aspects such as types of research output publication, direction of research output publication and volume of research output publication. The Nepali NLP works are broadly done on three categories namely Rule based, Machine Learning based and deep learning based [9].

First Nepali lexicon in various file formats with root word, head word, pronunciation, parts of speech for each word [10], spell checker and machine translation system introduced by Madan Puraskar Pustakalaya (MPP), Bhasa Sanchar or the NeLRaLEC project led by the MPP to develop annotated corpus Nepali National Corpus (NNC) are some of the NLP works in beginning time. The NNC corpus contains four different corpora written, spoken, parallel and speech corpus [11]. Nepali spell checker [10], Nepali grammer checker [12], Dobhase A machine translation system [13], online Nepali Dictionary [14], Nepali stemmer and morphological analyzer [15] are also other research work done in the Nepali NLP.

There has been least number of literature available in Natural Language processing (NLP) for Nepali language. The NLP's fundamental steps such as part of speech tagging [16], Morphological Analysis [15], Named-Entity Recognition [17] were analyzed with available resources.

Data Mining

The process of extracting or discovering meaningful information from raw data is called data mining. It extracts information from raw data and transform into understandable form for future use. It discovers the patterns using data analysis tool and techniques [18]. It is also known as knowledge discovery in data (KDD). It is the process of finding the pattern and information. It involve from data collection to visualization process for extracting information. It has mainly four steps as objective finding, data collection, applying data mining algorithm and evaluation of results.

2.1.1 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. Classification is basically four types namely binary, multi class, multi label and multi output [19].

Binary classification algorithm groups the data based on discrete or non-continuous values. It has usually two values. Logistic Regression and Support Vector Machine classifiers are strictly binary classifiers and can be used to perform multiclass classification with multiple binary classifiers. For example, email spam filtering sorts the email into either spam or ham class. Multiclass classification algorithm groups the data into more than two classes. Algorithms like SGD classifiers, Random Forest classifiers and naive Bayes classifiers can be used to handle multiple classes natively. For example, news classification categorizes different news into national, international, sports, education, business etc. classes.

The classification algorithm having two or more class labels and one or more class labels may be predicted for each example is known as multi label classification. Algorithms like Multi-label Decision Trees, Multi-label Random Forests, and Multi-label Gradient Boosting are used for multi-label classification. For example, classification of multiple objects like person, tree etc. in a photo. Multi output classification is a generalization of multi label classification where each label is multiclass i.e. it predicts multiple outputs. It gives two or more outputs after prediction. For example, model that classifies the fruit's color and type simultaneously.

2.1.2 Opinion Classification

Opinion is the subjective feelings presented in text by user. It is the expression of feelings of anger, disappointment, disgust, joy, happiness, amusement. Each opinion has a target. For example the opinion sentence "*This was the excellent movie I ever saw*" have the target "*movie*". The target is also known as an entity. Each entity has components, subparts, attribute and descriptors. In the sentence "*The ending plot was the superb part of the movie*" Here target of opinion is movie but the additional components is plot which is the end plot of the movie. These components and attributes are called aspects of the opinion [20].

In opinion analysis, various features of the text play important role in classification of opinion. The most important feature is terms used in the text. Terms includes words and phrases, but could also include punctuation, emoji, and emoticons (emotion icon i.e. representation of facial expression in pictorial form using characters) as those can also imply mood or feeling. The words that determine one's opinion or sentiments are called opinion words. Examples of opinion words are good, bad, hate, love, adore etc.

Opinion classification is the technique for classifying the opinion text sentences using different classification algorithm. This technique is useful to define the user view as opinion or not. An opinion helps the humans to carry out the decisions. As World Wide Web is increasing day by day, web documents can be seen as a new source of opinion for human beings [21]. In recent time, the amount of opinion by users is increasing rapidly. Users give their opinion in different platform such as online i.e. different social site, blog, online news portal and offline i.e. print media (newspaper).

The Opinion classification is done in various levels such as document level, sentence level and phrase level [22]. In document level, users view in document is classified. Each document acts as a single entity that has user's perspective on particular things only. For example, particular movie or product review. Sentence level classification is similar to document level and it only interact with particular sentences not the whole document. Sentence level opinion analysis is related with classification of subjectivity that determines the sentences into fact or opinion. Phrase level Analysis is done by identifying the phrases in sentence and classifying them in either of the class in this level. This research work is focused on analyzing opinion in Nepali text in sentence level.

The volume of opinion in Nepali language is also increasing rapidly. The 44.6% of total population used mother tongue as Nepali language and literacy rate is 65.9% [23]. In Nepal, the Registered Newspaper counts is 7,874 and registered online media is 2,760 [24]. The 82.8% population has the access of internet [25]. These data shows that the increase of people interaction on online platform. This helps people to give their view, opinion, and feeling on written text format in Nepali language. These may or may not be the opinion text. So classification should be done for these data.

2.1.3 Naïve 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,

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

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 "Naïve", 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.

Naïve bayes model have various types as Gaussian, Multinomial, Bernoulli etc. The Gaussian model assumes that features follow a normal distribution. If predictors use continuous values rather than discrete, model assumes these values are sampled from the Gaussian distribution.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where, σ = Standard deviation

μ = Mean

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. Bernoulli Model works similar to the

Multinomial classifier, but predictor variables are the independent Booleans variables i.e. if a particular word is present or not in a document. It considers a Bernoulli distribution of a random variable X.

$$P(X) = \begin{cases} p \; if \; X = 1 \\ q \; if \; X = 0 \end{cases}$$

where, q = 1-p; 0<p<1

Theoritical Model of Naïve Bayes Classifier

The probability model is a conditional model for a classifier over a dependent class variable C with a small number of outcomes or classes that are conditional on feature variables F_1 to F_n as [26];

$$p(C|F_1 \dots \dots \dots \dots \dots . F_n)$$

If the number of features is large, a model using probability tables is not feasible. Reformulating the model more tractable using Bayes Theorem as,

$$p(C|F_1 \dots \dots \dots \dots \dots . F_n) = \frac{p(C)p(F_1 \dots \dots \dots \dots \dots . F_n|C)}{p(F_1 \dots \dots \dots \dots \dots . F_n)}$$

Above equation also can be written as

$$posterior = \frac{prior * likelihood}{evidence}$$

The denominator of above equation does not depend on class variable C and features variables F_i i.e. denominator is constant. The numerator is equivalent to the joint probability model as,

$$p(C, F_1, \dots \dots \dots \dots \dots ., F_n)$$

Using the repeated applications of the definition of conditional probability, above equation can be written as,

$$p(C, F_1, \dots \dots \dots \dots \dots ., F_n)$$

$$= p(C)p(F_1 \dots \dots \dots \dots \dots . F_n|C)$$

$$= p(C)p(F_1|C)p(F_2 \dots \; F_n|C, F_1)$$

$$= p(C)p(F_1|C)p(F_2|C,F_1)p(F_3 \dots F_n|C,F_1,F_2)$$

$$= p(C)p(F_1|C)p(F_2|C,F_1)p(F_3|C,F_1,F_2)p(F_4 \dots F_n|C,F_1,F_2,F_3)$$

$$= p(C)p(F_1|C)p(F_2|C,F_1)p(F_3|C,F_1,F_2) \dots \dots p(F_n|C,F_1,F_2,F_3 \dots, F_{n-1})$$

In naive conditional independence, assume each feature F_i is conditionally independent of other feature F_j, for j ≠ i.

$$p(F_i|C,F_j) = p(F_i\,|C)$$

The joint model is;

$$p(C, F_1, \dots \dots \dots \dots \dots ., F_n)$$

$$= p(C)p(F_1|C)p(F_2|C)p(F_3|C) \dots \dots .\, p(F_n|C)$$

$$= \; p(C) \prod_{i=1}^{n} p(F_i|C)$$

From above assumptions, the conditional distribution over the class variable C can be written as;

$$p(C|F_1, \dots \dots ..F_n) = \frac{1}{Z}\, p(C) \prod_{i=1}^{n} p(F_i|C)$$

where Z = scaling factor dependent only on $F_1,\dots\dots,F_n$. (constant if values of feature variables are known)

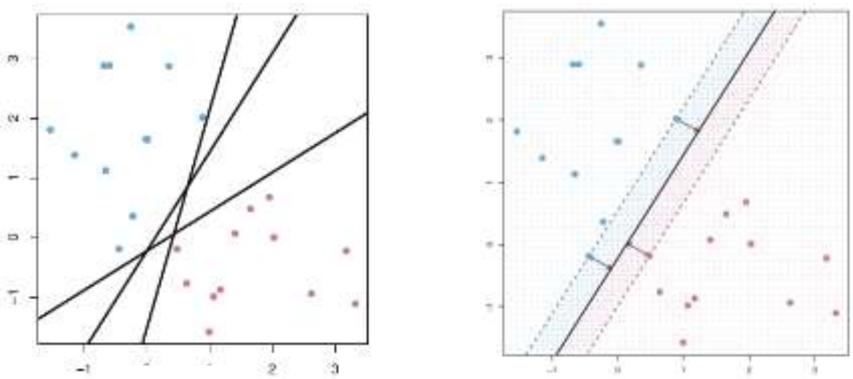
These types of models are manageable as they use prior probabilities of classes p(C) and independent probability distributions $p(F_i|C)$. If there are k classes for p(Fi) and model can be expressed in terms of r parameters, then the corresponding naive Bayes model has (k −1)+nrk parameters. Generally, k = 2 (binary classification) and r = 1 (Bernoulli variables as features) are common. The total number of parameters of the naive Bayes model is 2n + 1 (n is number of binary features used for prediction).

2.1.4 Support Vector Machine

It is Supervised Machine Learning algorithms that can be used for both classification and regression problems, mainly used for classification problems. If SVM is used for regression analysis it is known as support vector regression (SVR).

The main idea of SVM algorithm is to create the best line or decision boundary that separate the data points in different classes and helps to put new data point in correct class or category. SVM finds hyperplane using support vectors and margins.

The best decision boundary is called optimal hyperplane shown by solid line in figure 1(a) and other two dotted lines are hyperplane in figure 1(b). SVM chooses the extreme points that help in creating the hyperplane. These extreme points are called as support vectors. The distance between optimal hyperplane and hyperplane is known as margin which does not contains any data points.



a) b)

Figure 1 : a) Multiple hyperplane for linearly separable data b) The optimal hyperplane with maximum Margin **[27]**

Consider a linear classifier for a binary classification problem with labels y and features x. The class labels are denoted by y ∈ {-1,1}. The classifier is written as,

$$h_{w,b}(x) = g(w^T x + b)$$

where x is point vector, w is weight vector and b is bias. Here,

$$g(z) = 1 \text{ if } z \geq 0,$$

$$= -1 \text{ otherwise}$$

Given a training example (x $^{(i)}$, y $^{(i)}$), the functional margin of (w, b) with respect to the training example is

$$\gamma^{(i)} = y^{(i)}(w^T x + b)$$

The geometric margin with respect to training example is

$$\gamma^{(i)} = y^{(i)}\left(\left(\frac{w}{||w||}\right)^T x^i + \frac{b}{||w||}\right)$$

If ||w|| = 1, then the functional margin equals the geometric margin. So the optimization problem will be;

$$\min_{\gamma,w,b} \frac{1}{2}||w||^2$$

$$s.t.\ y^{(i)}(w^T x + b) \geq 1\ ; i = 1 \ldots..m$$

The constraints in above equation ensure that the maximum margin classifier is linearly separable data classifies each data points correctly.

The maximum or minimum of a multivariable function is finds with constraint using the Lagrange Multiplier.

The Lagrangian for optimization problem will be,

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i\ [y^{(i)}\left(w^T x^{(i)}\ + b\right) - 1]$$

where αi's is the Lagrange multipliers.

To find the maximum value, taking the derivative and set it to zero gives the following dual optimization problem.

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)} y^{(i)} \alpha_i \alpha_j < x^{(i)} x^{(i)} >$$

$$s.t. \alpha_i\ \geq 0, i = 1 \ldots \ldots m$$

$$\sum_{i=1}^{m} \alpha_i\, y^{(i)} = 0$$

By examining the dual form of the optimization problem, the resulting algorithm support vector machines, will be able to efficiently learn in very high dimensional spaces.

2.2 Literature Review

Shrestha and Bal implemented a named-entity based sentiment analysis framework on Nepali News Media Texts & visualized the popularity of the politicians via the time series graph of positive and negative sentiments. Experimental analysis showed Support Vector Machine had overall highest performance metrics for classifying the sentiments expressed in the sentences & found that Word2Vec with skip-gram was the optimal option for feature extraction [28].

Tamrakar, Bal and Thapa developed a model to detect the aspect based sentiment in Nepali text using Machine Learning classifier algorithms namely Support Vector Machine and Naïve Bayes. On various experiments showed that Bernoulli Naïve Bayes perform better than the SVM [29].

Regmi et.al. proposed a supervised machine learning based framework to make a distinction between 'facts' and 'opinions' in Nepali Subjective Text. A comparative analysis of 3 different supervised machine learning algorithms is done and support vector machine performs well even with a small size of dataset [30].

Gupta and Bal presented two main approaches for sentiment detection of Nepali texts. Lexical resource 'Bhavanakos' in which sentiment words are detected in Nepali texts to detect the sentiment in documents and machine learning based text classifier with annotated Nepali text data to classify the document. From the experiment done, the machine learning approach is better than resource based approach [31].

Pant and Yadav presented machine learning methods for detecting the sentiment expressed by movie reviews. Naive Bayes based machine learning technique is used for the classification of the sentiment. The results showed precision 79.23%, recall 78.57% and F-score 78.90% [32].

Thapa and Bal applied three Machine Learning classifiers, namely Support Vector Machine, Multinomial Naïve Bayes and Logistic Regression to classify book and movie reviews written in Nepali into 'Positive' and 'Negative' with 5-fold cross-validation techniques. From the experiment done, Multinomial Naive Bayes classifier performs higher accuracy than the other two classifiers [33].

Shahi and Yadav performed the mobile SMS spam filtering for Nepali text using Naïve Bayes and SVM and analyzed the accuracy. The model predicts the SMS as either spam or ham by using the content based filtering method. The empirical analysis showed that Naïve Bayes model have a better accuracy as compare to SVM model with accuracy 92.74% [34].

Paudel, Shahi and Sitaula proposed the hybrid feature extraction method for analyzing sentiment using syntactical and semantic information on publicly available Nepali COVID-19 tweets dataset called NepCov19Tweets. The dataset consists of Nepali tweets categorized into three classes Positive, Negative, and Neutral. They combined TF-IDF and FastText text representation methods for hybrid features & used nine machine learning classifiers such as Support Vector Machine, Naive Bayes, Decision Trees, K-Nearest Neighbor, Extreme Tree classifier, Logistic Regression, Random Forest, Multilayer Perceptron and AdaBoost. Based on the three feature representation methods such as TF-IDF, FastText, and Hybrid, experimental evaluation showed that hybrid feature extraction method outperforms other two individual feature extraction methods and also provided the better performance [35].

Acharya and Shakya studied the comparison of different classification algorithms such as SVM-RBF Kernel, SVM-Poly Kernel, NB Multinomial and Random Forest on Nepali text dataset. The comparison was done on basis of evaluation metrics Accuracy, Precision, Recall and F-Measure. On the basis of implementation output, SVM-Poly Kernel outperformed other three algorithms with Accuracy 82.76%, Precision 82.9%, Recall 82.8 % and F-Measure 82.7% [36].

Shahi and Pant proposed the automated news classification on Nepali text dataset. They evaluated Naive Bayes, SVM and Neural Networks algorithms on 4,964 documents with 20 different categories. The features are extracted with TF-IDF to test the models. The empirical results showed that the SVM with RBF kernel have outperformed other algorithms with the classification accuracy of 74.65% [37].

Dupakuntla et.al. classified the polarities of the reviews or opinions expressed in the Hindi language into positive or negative sentiments using a Naïve Bayes Classifier and evaluate the overall model’s performance with respect to various parameters. Also found that learning-based approaches for text sentiment analysis have a vast scope in developing an ideal sentiment classifier [38].

Ambasta proposed opinion classification system on twitter dataset into positive, negative and neutral category on the basis of different parameter like accuracy, precision, recall and F-Measure. Study found that the accuracy of Naïve Bayes classifier was much higher than the other two supervised learning algorithms i.e. Support Vector Machines and Maximum Entropy [39].

Arora et.al. studied Naive Bayes classifier and Support Vector Machine by analyzing and classifying sentiments of reviews (Movies, Amazon, Yelp) with different metrics like accuracy, precision, recall, roc score and f1-score. Experimental results showed that both algorithms have their own individual strengths and choice of the method should depend on the application and purpose [40].

Rahat, Kahir and Masum discussed Support vector machine and Naïve Bayes algorithm and compared accuracy, precession, recall value on airline review dataset. Experimental result showed that Support vector machine gave better result than Naïve Bayes algorithm [41].

Afzaal et.al. presented a fuzzy aspect based opinion classification system which efficiently extracts aspects from user opinions and perform near to accurate classification. Experimental results proved that the proposed system is effective in aspect extraction and also improve the classification accuracy [42].

Mikolov et.al. studied the quality of vector representations of words derived by various models on a collection of syntactic and semantic language tasks. They observed that it is possible to train high quality word vectors using very simple model architectures, compared to the popular neural network models. Experimental observation gave improvements in accuracy at much lower computational cost i.e., it takes less than a day to learn high quality word vectors from a 1.6 billion words data set [43].

Bouchlaghem, Elkhelifi and Faiz proposed a model that automatically classify opinions into positive, negative, neutral or non-opinion of twitter texts for Modern Standard Arabic (MSA). They preserved tweet particular features such as @usermentions, #hashtags and demonstrated that tweets specific features can improve classification performance. They applied Support Vector Machine (SVM), Naive Bayes (NB), J48 decision tree and Random forest algorithm on the tweets data. The experimental results showed that SVM gave the highest F measure 72%, and j48 classifier gave the highest precision 70,97% [44].

Phung et.al. studied supervised machine learning methods in opinion mining of online customer reviews on hotels in Vietnam on Agoda.com website. The study was used to find out which model is most compatible with the training dataset and apply that model to forecast opinions. The experimental results showed that Logistic Regression, Support Vector Machines and Neural Network methods have the best performance as compare to all six machine learning algorithm namely Naïve Bayes, Support Vector Machines, Logistic regression, Neural Network, Decision Tree, Random Forest [45].

Sawakoshi, Okada and Hashimoto proposed a method to classify sentences into either Opinion or Fact using SVM model on review data. A customer review influences other customer in travel business. They confirmed that if classification of Opinion and Fact sentences using SVM was better, there was not influence for classification of reviews into Positive and Negative by using only Opinion sentences [46].

CHAPTER 3

RESEARCH METHODOLOGY

This research study implement and analyzed the two supervised machine learning algorithm Naïve Bayes and SVM on Nepali opinion text dataset collected from different national newspaper's online portal. The analysis is done by comparing algorithms with different performance metrics namely accuracy, precision, recall and f1-score. The evaluation of performance of a particular scenario is known as performance analysis. The process include following steps (Figure 2).

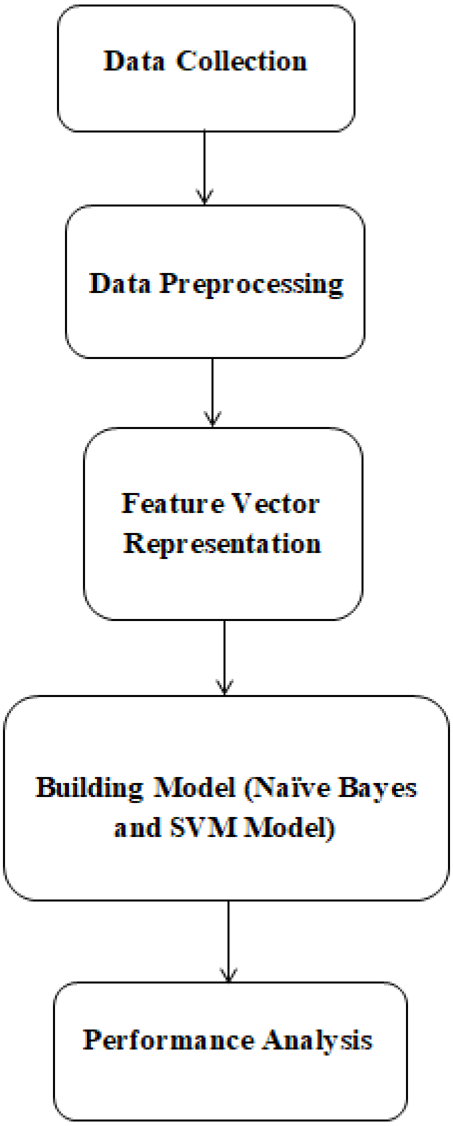

Data Collection
Data Preprocessing
Feature Vector Representation
Building Model (Naïve Bayes and SVM Model)
Performance Analysis

Figure 2 : Framework for Nepali Opinion Classification

3.1 Data Collection

The Nepali language is one of the script written in Devanagari, invented by Brahmins in 11th century [47]. Nepali is an Indo-Aryan language. It is the official language of Nepal. It is spoken throughout Nepal and other parts of the world. In Nepal, Nepali language is spoken by 78% of the population either as first or second language [23]. In worldwide, Nepali language is spoken by 17 million people. It consists of 36 consonants symbols, 13 vowel symbols and 10 numeral symbols along with different modifiers and half forms. The character set present in Nepali language is given (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, ऋ, ए, ऐ, ओ, औ, अं, अः

The Nepali sentences both opinion and non-opinion was collected from Nepali opinion section and other news section of various online news portals such as Gorkhapatra [48], Nagariknews [49], Setopati [50], Onlinekhabar [51]. The collected sentences were stored in comma separated file (CSV) format and labeled manually. The opinion and not opinion text in Nepali languages collected from different sources were used for classifying and analyzing performance of Naïve Bayes and SVM algorithm. The sample data was tabulated as (Table 2, Table 3).

Table 2 : Sample Nepali Opinion Sentences

Text
१. अझै पनि केही शंका छन् भने ती मेटिने गरी संसद्बबाट संकल्प प्रस्ताव पास गरेर जान पनि सकिन्छ।
२. रोजगारीबाट भएको कमाइ घर पठाएर स्वदेश फर्कन सक्ने वातावरण एवं वैदेशिक रोजगार क्षेत्रमा कुनै बदमास गरेमा कारबाही हुन्छ भन्ने मान्यताको विकास गराउनु आवश्यक छ ।
३. राष्ट्रसेवक कर्मचारीका लागि पनि कोभिड–१९ महामारीको अनुभवबाट एक किसिमको क्षमता र अनुभव प्राप्त भएको छ; जसले गर्दा अब हुने कुनै पनि महामारी नियन्त्रणमा अझ प्रभावकारी रूपमा प्रस्तुत हुन सकिने कुरामा विश्वस्त हुन सकिन्छ ।
४. यो अनुदान भएकाले सर्त योजना कार्यान्वयनमा मात्र सीमित रहनुपर्नेतर्फ सबैको ध्यान जान जरुरी छ ।
५. भ्रष्टाचार नियन्त्रण तथा निर्मूलीकरणका लागि राज्यले पनि भ्रष्टाचारसम्बन्धी नीतिको सशक्त एवं प्रभावकारी कार्यान्वयनका लागि तत्पर रहनु आवश्यक देखिन्छ ।

Table 3 : Sample Nepali Not Opinion Sentences

Text
१. डिजिटाइज हुन नसक्दा ती अखबार धुजाधुजा हुन थालिसकेका छन् ।
२. निर्वाचन कार्यालयले भने ७३ दलले पेस गरेको विवरणमाथि जाँचबुझ गरी अभिलेख राखी चैत ७ भित्र निर्वाचन आयोगमा पठाउनुपर्नेछ ।
३. सबस्टेसनमा १३२–३३ केभी, ३० एमभीए क्षमतार ३३–११ केभी, १६एमभिए क्षमताका एक ६.–एक वटा पावर ट्रान्सफर्मर रहेका छन् ।
४. राष्ट्रिय गौरवको आयोजनाका रूपमा रहेको जलविद्युत् आयोजनाको निरीक्षणपछि प्रधानमन्त्री देउवाले निर्देशन दिने कार्यक्रम छ ।
५. चन्द्रागिरि केबलकारको विद्युतीय आपूर्ति प्रणालीमा समस्या आएपछि केबलकारमा रमाउँदै गरेका यात्रु अचानक डेढ घण्टा थुनिए।

3.2 Preprocessing

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 stopword removal, tokenization, symbol and number removal and stemming.

i) Stop word 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 [10]. The stopword lists for English and other language are easily available but there is not any standard stopword 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, किनकि, पनि

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

ii) Tokenization

Tokens are the splitting unit in the string. Token is the character sequence like sentence, word which can be used for text analysis. Tokenization is the process of extracting the tokens. It is language specific task. In Nepali language, tokenization is performed mainly with words separation techniques because words are separated by the white spaces. The sentence level and word level tokenization can also be used.

iii) Symbol and Number removal

In text document, there are different types of symbols used to represent the information. They are not meaningful and not useful for text analysis. These are only used for organization of text in documents. These are removed before performing the machine learning model. 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

3.3 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, Fastext, ELMO 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.

i) 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;

$$Term\ Frequency\ (TF) = \frac{no.of\ times\ term\ occurrences\ in\ a\ document}{total\ number\ of\ words\ in\ a\ document}$$

ii) 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 [52].

Inverse Document Frequency (IDF)

$$= log_e \left\{ \frac{total\ number\ of\ documents}{number\ of\ documents\ which\ are\ having\ term} \right\}$$

Formula to calculate complete TF-IDF value is,

*TF-IDF = TF * IDF*

The step for TFIDF representation of text is [53].

Step 1: Identify unique words in the complete text data.

Step 2: For each sentence, create an array of zeros with the same length.

Step 3: For each word in each sentence, calculate the TF-IDF value and update the corresponding value in the vector of that sentence.

3.4 Classification

Classification is a Supervised Learning technique that is used to identify the category. In Classification, a program learns from the given data and then classifies new observation into a number of classes or groups. The Naïve Bayes and SVM classifier are used to classify the Nepali opinion text in this study.

3.5 Implementation

The experiment is performed in personal computer having Intel(R) Core (TM) i5-7200U CPU @ 2.50 GHz processor with 4 GB RAM and 64-bit Operating system. For implementation, Python programming language and Jupyter notebook as software environment is used to preprocess the data and implement the Naïve Bayes and SVM algorithm. 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 dissertation work are as follows:

1) **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.

2) **NumPy:** NumPy is a general purpose array processing Python library used for array variable and numerical computing.
3) **Matplotlib:** Matplotlib is a plotting library for 2D graphics in python programming language.
4) **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.
5) **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.

In this research work, two supervised machine learning algorithm was used for classification of Nepali opinion text data. The collected data were used for two methods- without preprocessing and preprocessing. The data was preprocessed with different NLP technique such as stop words removal, tokenization, removing number and punctuation. 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. In scikit learn, sklearn.model_selection module was used for train and test data. The data was splitted into training and testing dataset with 80% and 20% respectively. In Naïve Bayes method, multinomial Naïve Bayes was used which 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. For the zero probability case, Laplace smoothing was used. The whole process was performed with scikit learn library of machine learning.

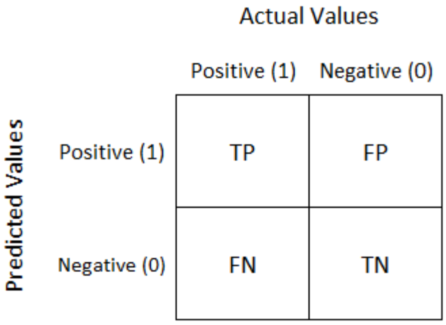
In SVM method, linear svm was used. In scikit learn, sklearn.svm.LinearSVC module was used for implementation of SVM model. SVM supports both dense and sparse input and the multiclass support is handled according to a one-vs-the-rest scheme. The performance measures were calculated from the confusion matrix which were calculated by sklearn.metrics module.

3.6 Performance Measure

The performance measure can be analyzed to check how well the model performs on the dataset. The performance was analyzed from various metrics such as accuracy, precision, recall and F1-measure. These can be calculated using the confusion matrix.

3.6.1 Confusion Matrix

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 [54]. One axis of the confusion matrix is predicted, and the other axis is the actual label.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 3 : Sample representation of Confusion Matrix

Here,

True Positives (TP): actual class of the data is True and the predicted is also True.

True Negatives (TN): actual class of the data is False and the predicted is also False.

False Positives (FP): actual class of the data is False and the predicted is True.

False Negatives (FN): actual class of the data is True and the predicted is False.

The above values from confusion matrix are used for determining the performance measure like accuracy, precision, recall, f-score of classifier.

3.6.2 Accuracy

Accuracy measures how classifier makes the correct prediction. It is the number of correct prediction divided by the total number of prediction. Accuracy gives better results when the target classes in the data are nearly balanced [19].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3.6.3 Precision

Precision measures how predictions are correctly identified. It is the ratio of correct positive predictions to the total number of positive predictions. The precision is depends on the value of prevalence. Prevalence is the ratio of actual true to total number of predictions.

$$Precision = \frac{TP}{TP + FP}$$

3.6.4 Recall

Recall measure proportion of actual positive to the proportion of negative that are correctly identified. It is the ratio of correct positive predictions to the overall number of positive examples in the test set.

$$Recall = \frac{TP}{TP + FN}$$

3.6.5 F1 score

F1-score is the combine single metric of precision and recall. It is a harmonic mean of precision and recall. It is focus on comparison of performance of two classifiers.

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

CHAPTER 4

RESULTS AND ANALYSIS

4.1 Data

The data for this research work were collected from different national newspaper's online portal such as Gorkhapatra, Onlinekhabar, Nagariknews, Setopati etc. The Nepali sentences both opinion and not opinion are collected and manually labeled them. Then data is in CSV file format. Data consists of opinion and not opinion Nepali sentences. The dataset contains 1,120 sentences labeled as opinion or not-opinion. Dataset contain 560 opinion sentences and 560 not-opinion sentences. The dataset is represented in Figure 4.

4.2 Performance of Naïve Bayes Algorithm

The experiment is done on the dataset with two methods. One is without preprocessed i.e. original collected data and other is with preprocessed data. The data is split into training and testing dataset with 80% and 20% respectively. The confusion matrix was created for test dataset using Naïve Bayes algorithm without preprocessed (Table 4).

Table 4 : Confusion Matrix for Naïve Bayes Algorithm without preprocessed dataset

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP) 91	False Positive (FP) 11
	Negative	False Negative (FN) 31	True Negative (TN) 91

Out of 224 instances 91 were found to be True Positive that represent model correctly predicts positive class i.e. opinion and also 91 instances were True Negative that represent model correctly predicts the negative class i.e. Not-opinion. There are 11 instances where model incorrectly predicts positive class as false positive and 31 instances as false negative that model incorrectly predicts negative class (Table 4).

The dataset was then preprocessed by removing stop words, symbols, numbers, and punctuation. Then data was split into training and testing dataset with 80% and 20% respectively. The confusion matrix was created for test dataset using Naïve Bayes algorithm on preprocessed data is given below.

Table 5 : Confusion Matrix for Naïve Bayes Algorithm on preprocessed dataset

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP) 92	False Positive (FP) 26
	Negative	False Negative (FN) 12	True Negative (TN) 94

Out of 224 instances 92 were found to be True Positive that represent model correctly predicts positive class i.e. opinion and also 94 instances were True Negative that represent model correctly predicts the negative class i.e. Not-opinion. There are 26 instances where model incorrectly predicts positive class as false positive and 12 instances as false negative that model incorrectly predicts negative class (Table 5).

4.3 Performance of Support Vector Machine Algorithm

The experiment is done same as Naïve Bayes algorithm with two methods i.e. without preprocessing and with preprocessing the dataset. The data was split into training and testing dataset with 80% and 20% respectively. The confusion matrix was created for test dataset using Support Vector Machine algorithm without preprocessed is given below.

Table 6 : Confusion Matrix for SVM Algorithm without preprocessed dataset

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP) 96	False Positive (FP) 14
	Negative	False Negative (FN) 26	True Negative (TN) 88

Out of 224 instances 96 were found to be True Positive that represent model correctly predicts positive class i.e. opinion and also 88 instances were True Negative that represent model correctly predicts the negative class i.e. Not-opinion. There are 14 instances where model incorrectly predicts positive class as false positive and 26 instances as false negative that model incorrectly predicts negative class (Table 6).

The dataset was then preprocessed and split into training and testing dataset with 80% and 20% respectively. The confusion matrix was created for test dataset using SVM algorithm on preprocessed data is given below.

Table 7 : Confusion Matrix for SVM Algorithm on preprocessed dataset

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP) 96	False Positive (FP) 12
	Negative	False Negative (FN) 22	True Negative (TN) 94

Out of 224 instances 96 were found to be True Positive that represent model correctly predicts positive class i.e. opinion and also 94 instances were True Negative that represent model correctly predicts the negative class i.e. Not-opinion. There are 12 instances where model incorrectly predicts positive class as false positive and 22 instances as false negative that model incorrectly predicts negative class (Table 7).

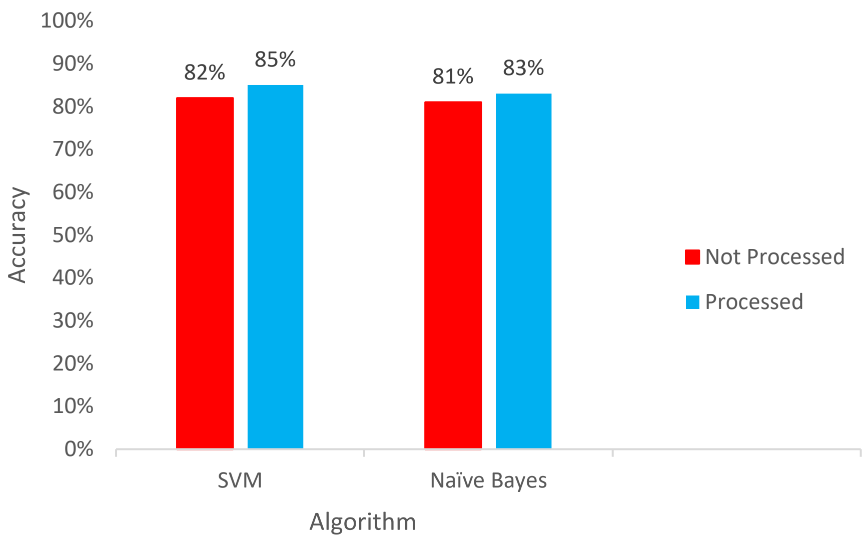
The performance of Naïve Bayes algorithm and SVM algorithm without preprocessing and preprocessing the dataset is calculated with different performance parameter and summarized (Table 8).

Table 8 : Comparative Results of Naïve Bayes and SVM

		Accuracy	Precision	Recall	F1-Score
Naïve Bayes Algorithm	Without Preprocessing	0.81	0.83	0.81	0.81
	Preprocessing	0.83	0.84	0.83	0.83
Support Vector Machine Algorithm	Without Preprocessing	0.82	0.83	0.82	0.82
	Preprocessing	0.85	0.85	0.85	0.85

The value of accuracy is increased after preprocessing the data over the not preprocessing the data. In Naïve Bayes model, accuracy is increase from 81% to 83% while in SVM model, the accuracy is increased from 83% to 85%. There is also increment of precision and recall value as compared to not preprocessing than the preprocessing the data. The F-measure is also increased (Table 8).

The graphical representation of accuracy for both classifiers that are used for the experiment is shown (Figure 5) and the comparative performance measure of both classifiers is shown (Figure 6).



.

Figure 4 : Comparison of Accuracy for Naïve Bayes and SVM Classifier

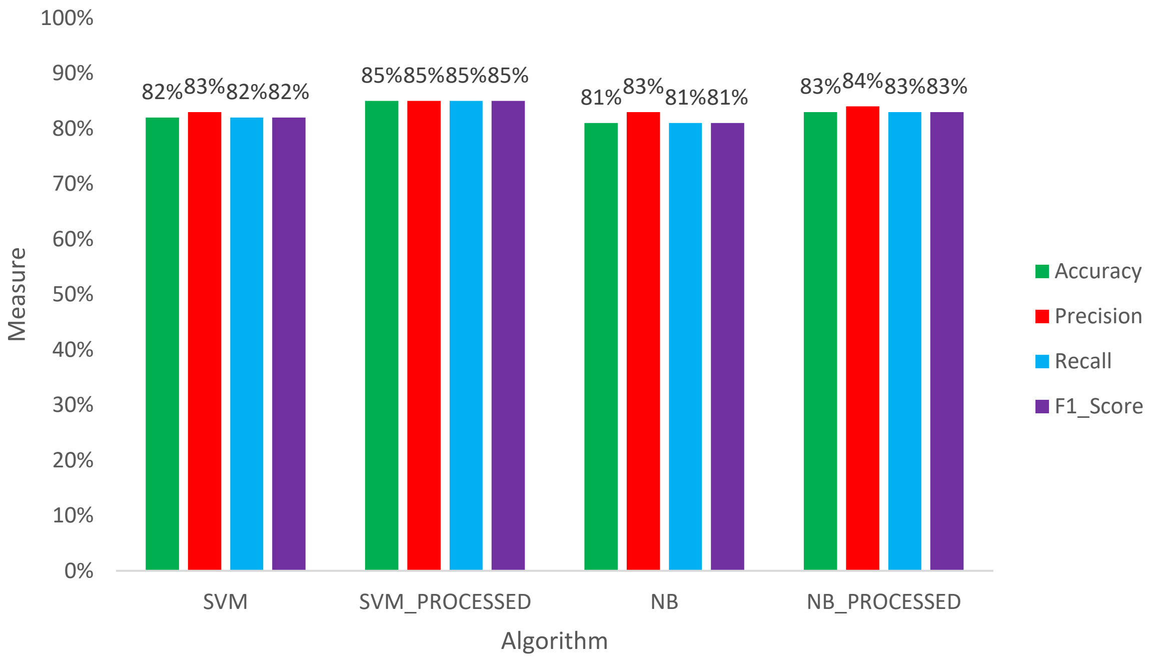


Figure 5 : Comparison of Performance Measure on Nepali Opinion Dataset

In implementing both classifiers with and without preprocessed data, SVM showed higher accuracy, precision and recall than the Naïve Bayes algorithm. In Naïve Bayes model, the accuracy is 81% and 83% on without processing data and processing data respectively. The precision and recall values were also increased from 83% to 84% and 81% to 83% respectively. F-measure is also increased from 81% to 83%.

In SVM model, the accuracy is increased from 82% to 85% while performing processing of the data. The precision and recall values were also increased from 83% to 85% and 82% to 85% respectively. F-measure is also increased from 82% to 85%.

In this research work, it is observed that Support Vector Machine (SVM) classifier showed higher accuracy, precision, recall and F1-score than Naïve Bayes classifier. So SVM is more efficient algorithm than Naïve Bayes for the classification of Nepali opinion text data.

CHAPTER 5

CONCLUSION AND FUTURE RECOMMENDATION

5.1 Conclusion

Performance analysis of two supervised machine learning algorithms Naïve Bayes and Support Vector Machine (SVM) were experimented for without preprocessing and preprocessing on Nepali text data in this study. The performance measures accuracy, precision, recall and F1-score were calculated and compared. The accuracy of SVM was 85% which is higher than accuracy of Naïve Bayes algorithm i.e. 83% on preprocessed the data. It concluded that the accuracy of both algorithms was improved after preprocessing as compared to without preprocessing the data. The study showed that the different performance measures precision, recall and f1-score were better for SVM as compared to Naïve Bayes algorithm. The Study concluded SVM model was the best model with higher values of performance metrics and is recommended for opinion classification of Nepali text data over the Naïve Bayes algorithm.

5.2 Future Recommendation

The performance analysis of Naïve Bayes and SVM algorithm gives the significant results in the Nepali opinion text dataset. In future, this technique can be used with different classifier for the better results. The use of stemmer in preprocessing step is added to analysis of the performance measure as of now there is no any standard stemmer is present for the Nepali language. The different variants of SVM and Naïve Bayes can be implementing for Nepali opinion classification and analysis will be done on the basis of performance metric.

REFERENCES

[1] "Merriam Webster," [Online]. Available: https://www.merriam-webster.com/dictionary/opinion. [Accessed 29 01 2022].

[2] V. Vapnik and A. Chervonenkis, "Theory of Pattern Recognition," 1974.

[3] S. Shalev Shwartz and S. Ben David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.

[4] K. Verspoor and K. Bretonnel Cohen, "Natural Language Processing," 2013.

[5] S. R. Joseph, H. Hlomani, K. Letsholo, F. Kaniwa and K. Sedimo, "Natural Language Processing: A Review," *International Journal of Research in Engineering and Applied Sciences,* vol. 6, no. 3, 2016.

[6] S. Bird, E. Klein and E. Loper, Natural Language Processing with Python, O’Reilly Media, Inc., 2009.

[7] K. W. Church and L. F. Rau, "Commercial applications of Natural Language Processing," *Communication of the ACM,* vol. 38, 1995.

[8] T. C. Rindflesch, "Natural Language Processing," *Annual Review of Applied Linguistics,* 1996.

[9] T. B. Shahi and C. Sitaula, "Natural Language Processing for Nepali text: a review," *Artificial Intelligence Review,* 2022.

[10] S. Bista , L. Khatiwada and B. Keshari , "Nepali lexicon development. PAN Localization," *Working Papers,* pp. 311-315, 2004.

[11] Y. Yadava , A. Hardie , R. R. Lohani , B. N. Regmi , S. Gurung , A. Gurung and T. McEnery , "Construction and annotation of corpus of contemporary Nepali Corpora," 2008.

[12] B. K. Bal and P. Shrestha, "Architectural and System Design of the Nepali grammer Checker, PAN Localization," *Working Paper,* 2007.

[13] S. Bista, B. Keshari, J. Bhatta and K. Parajuli, "Dhobhase: online English to Nepali machine translation system," in *26th annual conference of the Linguistic Society of Nepal,* 2005.

[14] B. K. Bal, "Towards building advanced natural language applications–an overview of the existing primary resources and applications in Nepali," in *Proceedings of the 7th workshop on Asian language resources (ALR7), Association for Computational*

Linguistics, Singapore, 2009.

[15] B. K. Bal and P. Shrestha, "A Morphological analyzer and a stemmer for Nepali, PAN Localization," *Working Papers,* 2004.

[16] T. B. Shahi, T. N. Dhamala and B. Balami, "Support vector machines based part of speech tagging for nepali text," *International Journal of Computer Applications,* vol. 70, 2013.

[17] S. B. Bam and T. B. Shahi, "Named entity recognition for nepali text using support vector machines," *Intelligent Information Management,* vol. 6, p. 21, 2014.

[18] J. Han and M. K. Kamber, Data Mining: Concept and Techniques, Morgan Kaufmann Publishers, 2012.

[19] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly Media, Inc, 2019.

[20] M. Squire, Mastering Data Mining with Python - Find patterns hidden in your data, 2016: Packt Publishing Ltd..

[21] R. Sharma, S. Nigam and R. Jain, "Opinion Mining In Hindi Language: A Survey," *International Journal in Foundations of Computer Science & Technology,* vol. 4, 2014.

[22] L. B. Thapa and B. K. Bal, "Classifying sentiments in Nepali subjective texts," *7th International Conference on Information, Intelligence, Systems & Applications (IISA),* 2016.

[23] "National Population and Housing Census 2011," Central Bureau of Statistics Government of Nepal , 2014.

[24] "46th Annual Report 2077/78," Press Council Nepal, 2078 B.S..

[25] "Economic Survey 2077/78," Ministry of Finance, Government of Nepal , 2078 B.S..

[26] M. Narasimha Murty and V. Susheela Devi, Pattern Recognition, An Algorithmic Approach, Springer, Universities Press (India) Pvt. Ltd., 2011.

[27] G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013.

[28] B. B. Shrestha and B. . K. Bal, "Named-Entity Based Sentiment Analysis of Nepali News Media Texts," in *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications,* Suzhou, China, 2020.

[29] S. Tamrakar, B. K. Bal and R. B. Thapa, "Aspect Based Sentiment Analysis Of Nepali Text Using Support Vector Machine And Naive Bayes," *Technical Journal Nepal Engineers' Association, Gandaki Province*, vol. 2, 2020.

[30] S. Regmi, B. K. Bal and M. Kultsova, "Analyzing Facts and Opinions in Nepali Subjective Texts," in *8th International conference on information, intelligence, systems & applications (IISA)*, 2017.

[31] C. P. Gupta and B. K. Bal, "A Bootstrap Approach for Sentiment Analysis of texts in the Nepali Language," in *In Proceedings of the Cognitive Computing and Information Processing (CCIP),* Nioda, India, 2015.

[32] A. K. Pant and A. Yadav, "Sentiment Analysis on Nepali Movie Reviews using Machine Learning," 2014.

[33] L. B. R. Thapa and B. K. Bal, "Classifying Sentiments in Nepali Subjective Texts Classifying sentiments in Nepali subjective texts," in *7th International conference on information, intelligence, systems & applications (IISA)*, 2016.

[34] T. B. Shahi and A. Yadav, "Mobile SMS Spam Filtering for Nepali Text Using Naïve Bayesian and Support Vector Machine," *International Journal of Intelligence Science,* 2014.

[35] T. B. Shahi, C. Sitaula and N. Paudel, "A Hybrid Feature Extraction Method for Nepali COVID-19-Related Tweets Classification," *Hindawi Computational Intelligence and Neuroscience,* 2022.

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

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

[38] V. P. Dupakuntla, H. Veeraboina, M. V. K. Reddy, M. M. Satyanarayana and Y. S. Sameer, "Learning Based Approach for Hindi Text Sentiment Analysis Using Naive Bayes Classifier," *International Journal of Innovations in Engineering Research and Technology,* vol. 7, no. 8, 2020.

[39] S. Ambasta, "Opinion Classification System Using Supervised Learning Algorithm," *International Journal of Current Research* , vol. 8, no. 10, 2016.

[40] A. Arora, P. Patel, S. Shaikh and A. Hatekar, "Support Vector Machine versus Naive

Bayes Classifier: A Juxtaposition of Two Machine Learning Algorithms for Sentiment Analysis," *International Research Journal of Engineering and Technology*, vol. 07, no. 07, 2020.

[41] A. M. Rahat, A. Kahir and A. K. M. Masum , "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," in *8th International Conference on System Modeling & Advancement in Research Trends*, 2019.

[42] M. Afzaal, . M. Usman, A. C. M. Fong, S. Fong and Y. Zhuang, "Fuzzy Aspect Based Opinion Classification System for Mining Tourist Reviews," *Advances in Fuzzy Systems*, 2016.

[43] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representation in Vector Space," in *International Conference on Learning Representations*, 2013.

[44] R. Bouchlaghem, A. Elkhelifi and R. Faiz, "SVM based approach for opinion classification in Arabic written tweets," in *ACS 12th International Conference of Computer Systems and Applications (AICCSA) IEEE*, 2015.

[45] T. K. Phung, N. An Te and T. T. T. Ha, "A machine learning approach for opinion mining online customer reviews," in *21st ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter)*, 2021.

[46] Y. Sawakoshi, M. Okada and K. Hashimoto, "An Investigation of Effectiveness of “Opinion” and “Fact” sentences for Sentiment Analysis of Coustomer reviews," in *International Conference on Computer Application Technologies*, 2015.

[47] R. P. Tripathi, F. S. Verma and M. L. Sriwasthav, Hindi Vishwa Kosh, Nagari Pracharini Sabha, Varanasi, 1966.

[48] "Gorkhapatra," [Online]. Available: https://gorkhapatraonline.com/. [Accessed March 2022].

[49] "Nagarik News," [Online]. Available: https://nagariknews.nagariknetwork.com/. [Accessed March 2022].

[50] "Setopati," [Online]. Available: https://www.setopati.com/. [Accessed March 2022].

[51] "Onlinekhabar," [Online]. Available: https://www.onlinekhabar.com/. [Accessed March 2022].

[52] C. D. Manning, P. Raghavan and H. Schutze, Introduction to Information Retrieval,

CAMBRIDGE UNIVERSITY PRESS, 2008.

[53] K. Bhanot, "towardsdatascience," April 2022. [Online]. Available: https://towardsdatascience.com/different-techniques-to-represent-words-as-vectors-word-embeddings-3e4b9ab7ceb4.

[54] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management,* vol. 45, p. 427–437, 2009.

[55] [Online]. Available: https://medium.com/@anishajain2910/confusion-matrix-30249214041d. [Accessed 31 01 2022].

[56] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representation in Vector Space," in *Proceedings of the International Conference on Learning Representations* , 2013.

[57] K. Bhanot, "towardsdatascience," [Online]. Available: https://towardsdatascience.com/different-techniques-to-represent-words-as-vectors-word-embeddings-3e4b9ab7ceb4. [Accessed May 2021].

[58] F. Aiolli, R. Cardin and A. Sperduti, "Preferential text classification: Learning algorithm and evaluation measures," *FRCIM News,* 2009.