A Major Project Final Report on

Nepali News Classifier

Submitted in Partial Fulfillment of the Requirements for
the Degree of **Bachelors' of Engineering in Computer Engineering**
under Pokhara University

Submitted by:
Sanjeev Poudel, 15374
Anil Thapa, 15302

Under the Supervision of
Asst. Prof. Kamal Chapagain

Date:
25 Nov 2019

Department of Computer Engineering
NEPAL COLLEGE OF INFORMATION TECHNOLOGY

Balkumari, Lalitpur, Nepal

ACKNOWLEDGEMENTS

This project report is the leading way to the better understanding and implementing the acquired knowledge in the practical field. We take this opportunity to express our profound gratitude and deep regards to all those who have directly and indirectly helped us to complete this Project.

We would like to express our sincere gratitude to, our project supervisor Asst. Prof. Kamal Chapagain, for the constant guidance with his insightful ideas and valuable suggestions and encouragement during the period of the project.

We express our gratitude to Er. Resha Deo, the Head of Department of Computer Engineering, NCIT and to Asst. Prof. Dr. Roshan Chitrakar from the Project and Research Department, NCIT for providing a base for developing the major project of our own that professionalize our knowledge and provides us a remarkable experience for our career.

A special thanks goes to our teachers, friends and family for continuous support, encouragement and motivation during the project. Last but not the least we are grateful to everyone for their direct and indirect support in the progress of completing this project.

ABSTRACT

This Project is about News Classification which is one of the most popular applications of Natural Language Processing (NLP). In many real-world scenarios, the ability to classify documents into a fixed set of categories is highly desirable. Common scenarios include classifying a large amount of unclassified archival documents such as newspaper articles, legal records and academic papers. The fundamental task for this purpose is to collect data to train the models. The dataset is prepared from web scrapping technique, collecting news articles from various Nepali news portals. Datasets are categorized in 12 different groups. These groups of datasets are used for training the models. Different Classification algorithms are used for e.g. Support Vector Machine, Naïve Bayes etc. We implement the trained models in GUI using Tkinter Library of Python, as Desktop Application. While evaluating the performance metrices of the models, the Logistic Regression Model gives better accuracy than other models. This project also deals with some basic NLP related tasks such as tokenization, POS tagging, Romanization and Transliteration of Nepali Language.

Keywords: *GUI, NLP, POS tagging, Scikit Learn, Tkinter, Web Scrapping*

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

1.1. Background

Natural Language Processing is special domain interrelated with Computational Linguistics, Probabilities and Statistics and the Computer Science, especially Artificial Intelligence [1] Natural Language Processing deals with developing systems intelligence enough to have ability of understanding natural human languages, processing them and generating responses in natural language in real time [2].

This Project is about News Classification or Text Classification which is one of the most popular applications in NLP. Amazing development of Internet and digital library has triggered a lot of research areas. Text categorization is one of them. Text categorization is a process that group text documents into one or more predefined categories based on their contents. In many real-world scenarios, the ability to classify documents into a fixed set of categories is highly desirable. Common scenarios include classifying a large amount of unclassified archival documents such as newspaper articles, legal records and academic papers [3]. For example, newspaper articles can be classified as ‘sports’, ‘health’, ‘politics’ or ‘entertainment’.

This project looks specifically at the task of taking any uncategorized article from the user and classify using our trained models with the probability or accuracy of prediction. We have used several Classification Algorithms to train the models, and we compare the prediction accuracy of each of these algorithms used. Since our models are trained from already classified documents so we are able to make use of supervised classification techniques [4]. This project focuses at investigating and implementing techniques which can be used to perform automatic article classification for this purpose. We randomly split this archive of classified documents into training and testing sets for our classification systems. This project experiments with different natural language feature sets as well as different statistical techniques using these feature sets and compares the performance in each case.

This document provides the scope and context of the project that has been undertaken. It also provides a schedule followed and task performed for the completion of the project, including a list of all the deliverables and presentations required.

As an emerging area of Modern Science, NLP is being developed as an advancing and interconnecting tool for numerous Languages existing in this world. There are thousands of Languages existing in the world. They all have their own script and grammars to govern the spoken and text formats of communication. Nepali is a Language of Indo-Aryan family written in Devanagari script, spoken by about 32 million people specially on the state of Nepal along with Bhutan, Indian states of Sikkim, WB, and also in Burma. In spite of being a rich Language in, Technological tasks has not been done yet. Though Madan Puraskar Pustakalaya and Nepal Bhasha Sanchar Project has initiated some tasks in the sector of Nepali Language Processing back in 2006 [5]. The tasks have not been advanced as expected. In comparison to contemporary Languages in the world, we are still lagged backward. So, we initiated this project in order to study the status of Nepali Language Processing and to understand the sectors on which we can contribute.

1.2. Problem Statement

The text classification problem is an Artificial Intelligence research topic, especially given the vast number of documents available in the form of web pages and other electronic texts like emails, research papers, news articles, discussion forum postings and other electronic documents. Different supervised learning algorithms are available for text classification, and our problem is to investigate which supervised machine learning methods are best suited for news classification problem. On observing the accuracy of prediction, we need to obtain best model for the classification problems. The classifier makes the assumption that each new complaint is assigned to one and only one category. This is multi-class text classification problem [6]. Since, there are not any classification application available so far for Nepali News or Documents, obtaining the best models for 'Nepali News Classification' is really a challenging problem to be considered. In context of not getting appropriate application for such purposes, to develop the news categorizer/classifier along with other NLP features in Desktop Application Format is also the difficult target for this project. This project is focused in comparing among different classifiers and obtain the best classifier model for Nepali News.

1.3. Project Objectives

The general objective of Nepali News Classification Project is to develop a Nepali News Classifier application based on GUI, and implementing various tools and techniques of NLP.

The specific objectives of this project work are as follows:

- To analyze and Compare different models for News Classification
- To ease the procedure of Classifying Uncategorized or Unclassified News into given categories.
- To automate the news post categories in papers and newsrooms
- To reduce the subjective mess-up of news
- To Predict the possible class of input text/news articles with the score of accuracy.
- Implementing some basic NLP tasks integrated in a single application

1.4. Significance of the study

The importance of this project is that it is based on research on Machine Learning, different types of learning algorithms, and Nepali Linguistic. The news/ text classification problem is an Artificial Intelligence research topic, especially given the vast number of documents available in the form of web pages and other electronic texts like emails, discussion forum postings and other electronic documents. User can categorize the text document or news articles. Comparative analysis of performance of popular Classification Algorithm This has significant importance in case of Nepali Natural Language Processing. It can be further expanded to explore other different applicability in the sector of Nepali Language Processing, which still has many spectrums to be uncovered through such researches. The findings of this study provide students and scholars an efficient way and motivate to further advancing in Nepali NLP tasks.

1.5. Scope and Limitations

The scope of the project is to implement techniques of natural languages processing. To meet the requirements, we use different approach of document classification. Any unknown or untrained text document or news article are provided to be classified to the model. It will predict with accuracy the probability of that input text to fall under the available or pre-set categories. This project of text classification plays an important role and on further research and works on it, may be remarkably applicable in future for other problems of Nepali NLP like-

- Newspapers/ Media House can make use of it
- NLP Researchers and Students can further work on it for advancing Nepali NLP field
- Information extraction
- Summarization
- Text retrieval
- Question-answering
- Product Review Analytics
- Social Media Sentiment Analysis
- Nepali Movie Review Analysis etc.

The limitation of this application is that can’t classify to the class which has not been pre-specified and trained the models with. User can not define their own custom classes for their inputs to be categorized.

2. LITERATURE REVIEW

Natural Language Processing has been one of the most investigated research topics since the decades. One of the most popular applications of machine learning is the analysis of categorical data, specifically text data. Issue is that, there are a ton of resources out there for numeric data but very little for texts. The top technological firm such as the Google, Microsoft, Facebook and others have invested much dollar in the research of NLP. The Google has many products such as Google Voice, Google Translator, Google Input Tools which are available in many languages [7]. Among these products, most of them include Nepali Language but not as comparable to other language like English and European Language. However, the research on Nepali language processing is in its fledgling stage i.e. there are many rooms to fill out yet.

Primarily, most of the works on text classification were based on English language. Apart from English language, some form of text classification system exists for European languages such as Italian, German, Spanish, etc. and Asian languages such as Arabic, Chinese and Japanese. Nepali is a morphologically rich language that has a fairly complicated orthography. Due to this, many language features have to be taken into consideration to build an efficient text classification model. Even so, commendable efforts have been made in the field of Nepali text classification using various methods.

2.1. Reviews of Existing Works

This topic basically discusses the work done by various authors, students and researchers in brief in the area under discussion, which is News Classification using Machine Learning algorithms. The purpose of this section is to critically summarize the current knowledge in the field of document classification.

Classification of Text Documents

In the work [8] done by Y. H. LI AND A. K. JAIN they performed document classification on the seven class Yahoo newsgroup data set. The data set contained documents divided into following classes: International, Politics, Sports, Business, Entertainment, Health, and Technology. They employed Naïve Bayes, Decision Trees, Nearest neighbor classifier and the Subspace method for classification. They also performed classification using the combination of these algorithms. They adopted the commonly used bag of words document representation scheme for feature representation.

In this paper, they used a Binary representation for Naïve Bayes and Decision trees method. Whereas, they used Frequency representation in Nearest neighbor classifier and the Subspace method classifier to calculate the weight of each term.

In their experiments, all the four machine learning algorithms performed well. The Naïve Bayes gave the highest accuracy for first test data set but was outperformed by subspace method for the second test data set.

Support Vector Machines for Text Categorization

In the paper [9], the authors A. Basu, C. Watters, and M. Shepherd compared support vector machine with an artificial neural network for the purpose of text classification of news items. The authors used Reuters News data set for their comparative study. As the name suggests, theReuters-21578 dataset contains a collection of 21,578 news items that are divided across 118 categories. These are set of binary classification algorithms proposed by Vapnik. It works by finding a hyperplane that separates the two classes with maximum margin. SVM can operate with a large feature set without much feature reduction. This makes SVM an accomplished algorithm for classification.

After experiments, the authors concluded that SVM performed much better than Artificial Neural Network for both IQ87 and IQ57. Since SVM is also less computationally expensive, the authors recommended SVM over ANN for data set containing fewer categories with short documents.

Text Categorization with SVM: Learning with Many Relevant Features

In this paper [10] the author Thorsten Joachims explored and identified the benefits of Support Vector Machines (SVMs) for text categorization. The author performed stemming as part of preprocessing before creating the feature vectors. For generating feature vectors, the authors made use of word counts. Thus, each document was represented as vector of integers where each integer represented the number of times a corresponding word occurred in the document. To avoid large feature vectors the author only considered those words as features that took place more than three times in the document. The authors also made sure to eliminate stop words in making feature vectors. This representation scheme still led to very high-dimensional feature spaces containing 10000 dimensions and more. To reduce the number of features and overfitting, information gain criterion was used. Thus, only a subset of features was selected based on Information gain. The data set author used was ModApte split of theReuters-21578 dataset which is compiled by David Lewis. This dataset contained 9603 training documents and 3299 test documents. The dataset contained 135 categories of which only 90 were used since only 90 categories had at least one training and test sample. The author compared the performance of SVMs with Naïve Bayes, Rocchio, C4.5, and KNN for text categorization.

The author concludes that, among the conventional methods KNN performed the best on Reuters data set. On the other hand, SVM achieved the best classification results and outperformed all the conventional methods by a good margin. The author states that SVM can perform well in high dimensional space and thus does not mandate feature selection which is almost always required by other methods. The author also concludes that SVMs are quite robust and performed well in virtually all experiments.

Nepali News Classification using Naïve Bayes and Support Vector Machines

This research [11] evaluates some most widely used machine learning techniques, mainly Naive Bayes, SVM and Neural Networks, for automatic Nepali news classification problem. To experiment the system, a self-created Nepali News Corpus with 20 different categories and total 4964 documents, collected by crawling different online national news portals, is used. TF-IDF based features are extracted from the preprocessed documents to train and test the models. The average empirical results show that the SVM with RBF kernel is outperforming the other three algorithms with the classification accuracy of 74.65%. Then follows the linear SVM with accuracy 74.62%, Multilayer Perceptron Neural Networks with accuracy 72.99% and the Naive Bayes with accuracy 68.31%.

Some other works worthy to mention here are Shahi and Yadav analyses the effects of two classification techniques, the Naive Bayes and SVM, to develop a Mobile SMS spam filtering for Nepali text [12]. Dangol and Timalsina implement various Nepali language specific features such as filtering stop-words, word replacements and removal of word suffices using Nepali language morphology to reduce the number of dimensions in Vector Space Model [13].Similarly, Bam and Shahi classify rigid designators in Nepali text such as proper names, biological species and temporal expressions into some predefined categories which plays an important role in different fields such as Machine translation, Information Extraction, Question Answering System, etc. [14].

2.2. Overview of News Categorization

With the increasing of information on the internet and development of digital articles, people urgently need an efficient tool to automatically classify the information into categories. In this way, we can easily search, filter and store the large amount of resources. In many real-world scenarios, the ability to automatically classify documents into a fixed set of categories is highly desirable. Other scenarios involve classifying of documents as they are created. Examples include classifying movie review articles into ’positive’ or ‘negative’ reviews or classifying only blog entries using a fixed set of labels Natural language processing others powerful techniques for automatically classifying documents. These techniques are predicated on the hypothesis that documents in different categories distinguish themselves by features of the natural language contained in each document. Salient features for document classification may include word structure, word frequency, and natural language structure in each document. Automated text categorization is a process that assigning pre-defined category labels to new documents based on the contents.

Text categorization has many applications. For example, we can classify web pages into different categories to speed up the internet search, which is very useful for some search engines like Yahoo. Text categorization can be applied to filter emails to judge if it is spam email and further folder the emails. For news agencies, such as Globe and Mail, they receive thousands of articles a day. Articles can be classified to several categories like sports, politics, medical and etc. by text categorization methods. In digital library, people use key words to index articles, text categorization can also be used to classify the digital articles according to the subjects or key words. Usually there are two stages involved in text categorization, training stage and testing stage. In training stage, we need a learning algorithm. Documents are preprocessed and are trained by a learning algorithm to build the classifier. In testing stage, classifier is validated and used to categorize documents.

3. METHODOLOGY

We have worked, following these methodologies for the application of knowledge, skills, tools and techniques to a broad range of activities in order to meet the requirements of this project.

3.1. Technical description of the Project

Text Classification is an example of supervised machine learning task since a labelled dataset containing text documents and their labels is used for train a classifier. An end-to-end text classification pipeline is composed of these main components:

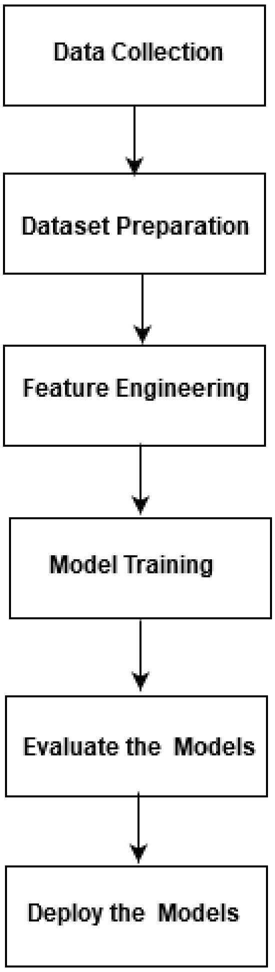


Figure 1 General Task Flow Chart

1. **Data Collection:** The very first step is the data collection. In this step, we collected news articles data from sources like – newspapers' websites and online news portals (namely: www.setopati.com, www.newsagro.com, www.edukhabar.com,

www.ictnews.com, www.filmykhabar.com, www.nepalkhabar.com, www.karobardaily.com, www.sahityasangraha.com, through the technique of web scraping[14] and stored in categorical folders in text format.

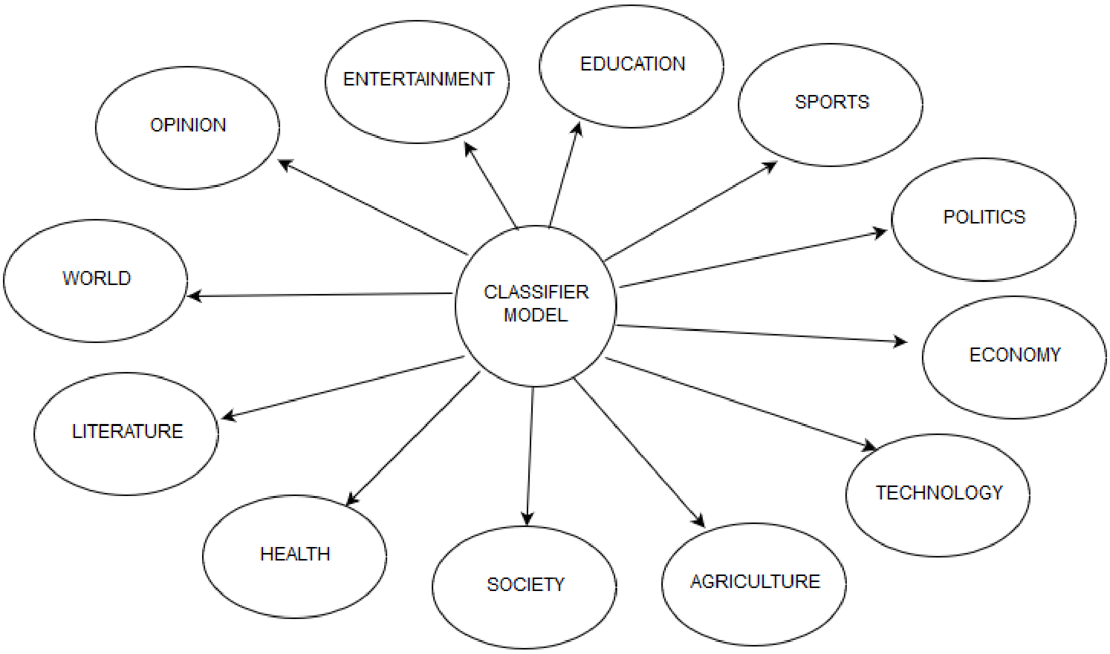


Figure 2 Categories predefined for the project

2. **Dataset Preparation:** The next important step is the Dataset Preparation step which includes the process of loading a dataset and performing basic pre-processing. We loaded each categorical text files into excel with labeled category of each. We have 12 predefined classes for this project. Stopwords removal, punctuation and other unnecessary contents from the news data was done during the preprocessing.

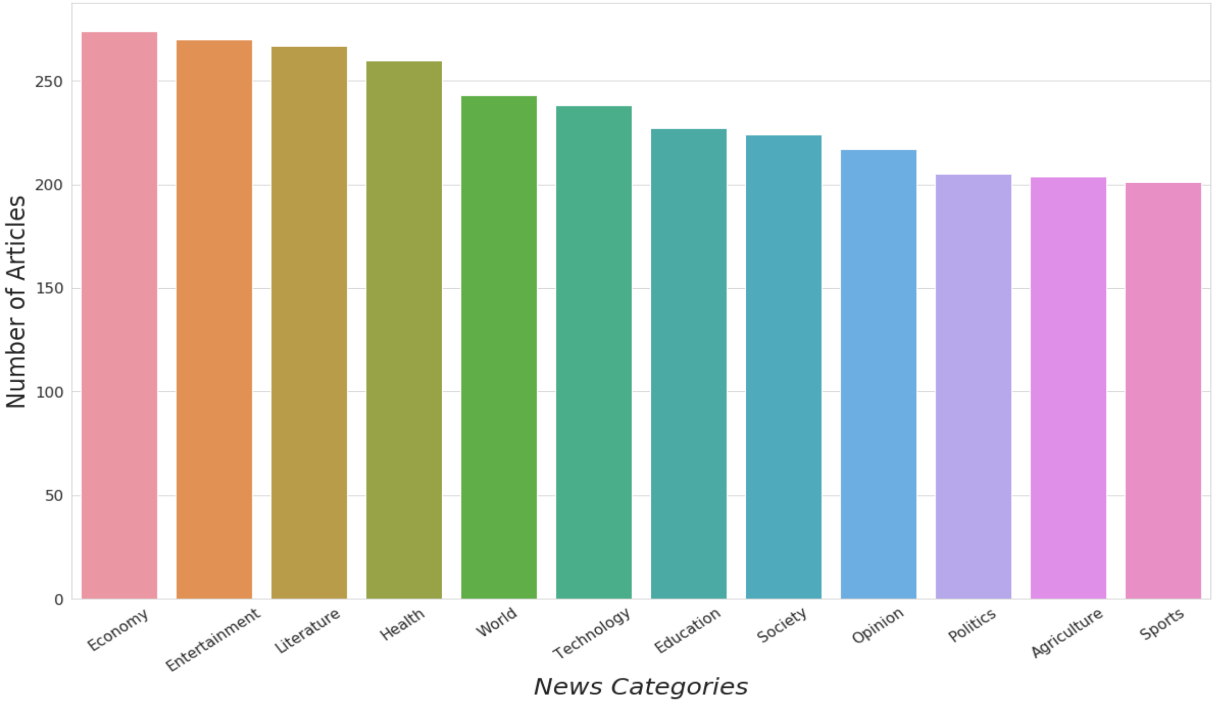


Figure 3 Categorical News Numbers used for training

Figure 3 shows the numbers of news articles and their categorical distribution, that are used for the training of the models.

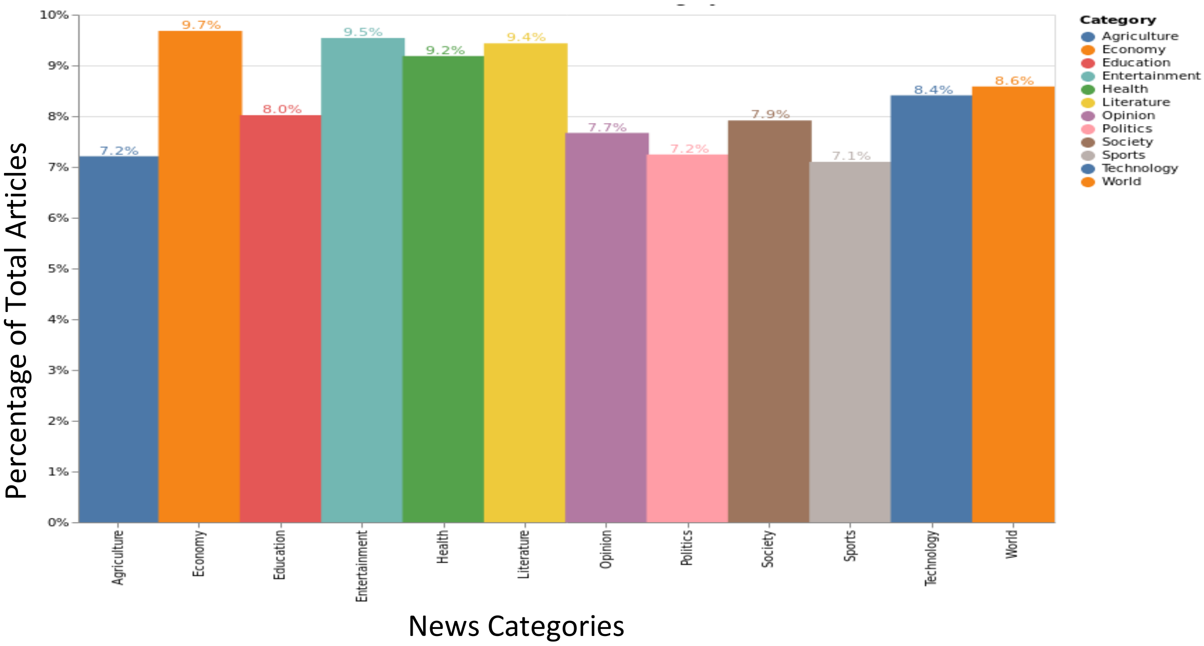


Figure 4 Percentage Distribution of Categorical News Articles

Figure 4 shows how the training data sets are formed with the composition of different categories. Economy category has most numbers of articles in the training data sets with 9.7% of total articles. Then, Entertainment, Literature and Health contribute highest percentage of articles in the data sets used for training.

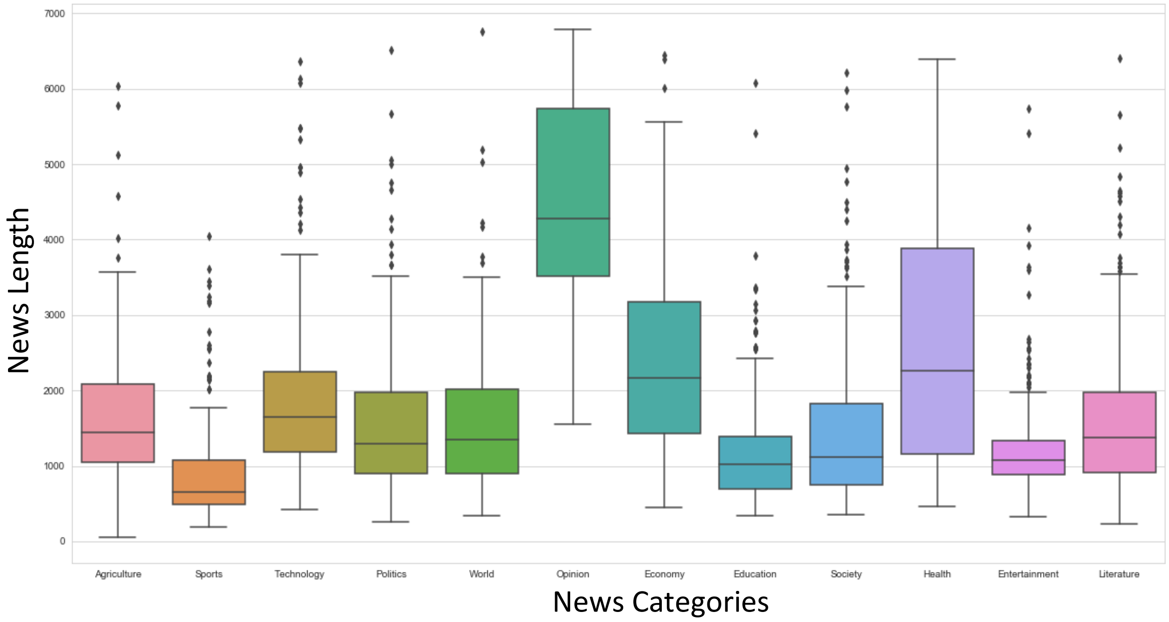


Figure 5 Box Plot for Length Distribution of News

Figure 5 showing the box plot illustrates the variation of length of training news articles, in different categories. It shows that most of the articles from opinion category have larger length varying from 3000 to 8000 words. Then, Health and Economy Category also have longer news articles for training, than other categories.

3. **Feature Engineering:** The next step is the Feature Engineering in which the raw dataset is transformed into flat features which can be used in a machine learning model. This step also includes the process of creating new features from the existing data. The dataset is then splitted into train and test sets in the ratio of 80:20.

4. **Model Training:** The final step is the Model Building step in which a machine learning model is trained on a labelled dataset. The hyperparameters are set depending on the algorithms. Cross Validation and Grid Search Techniques are used for the training of each models. We trained our data with six different types of supervised Algorithms (namely: SVM, k-NN, Logistic Regression, Random Forest, Naïve Bayes and Gradient Boosting). The models are pickled to be used or deployed in the application we built.

5. **Evaluating the Algorithms/ Test the Trained Models**: The last and final step of solving a machine learning problem is to evaluate the performance of the algorithm.

Different metrics are used to evaluate an algorithm. We analyze the Classification Report, Confusion Matrix and Accuracy Score Metrices and compare these of all the models. The train set accuracy is obtained very high of about 99 % in each but discarding the models with high bias and overfitting, the Logistic Regression Classifier Model is found to have highest accuracy of about 79.32 % in the test set.

6. **Deployment of News Classifier Model:** We finally, implement the best models obtained in a Tkinter [15] based GUI application.

3.2. Use Case Diagram

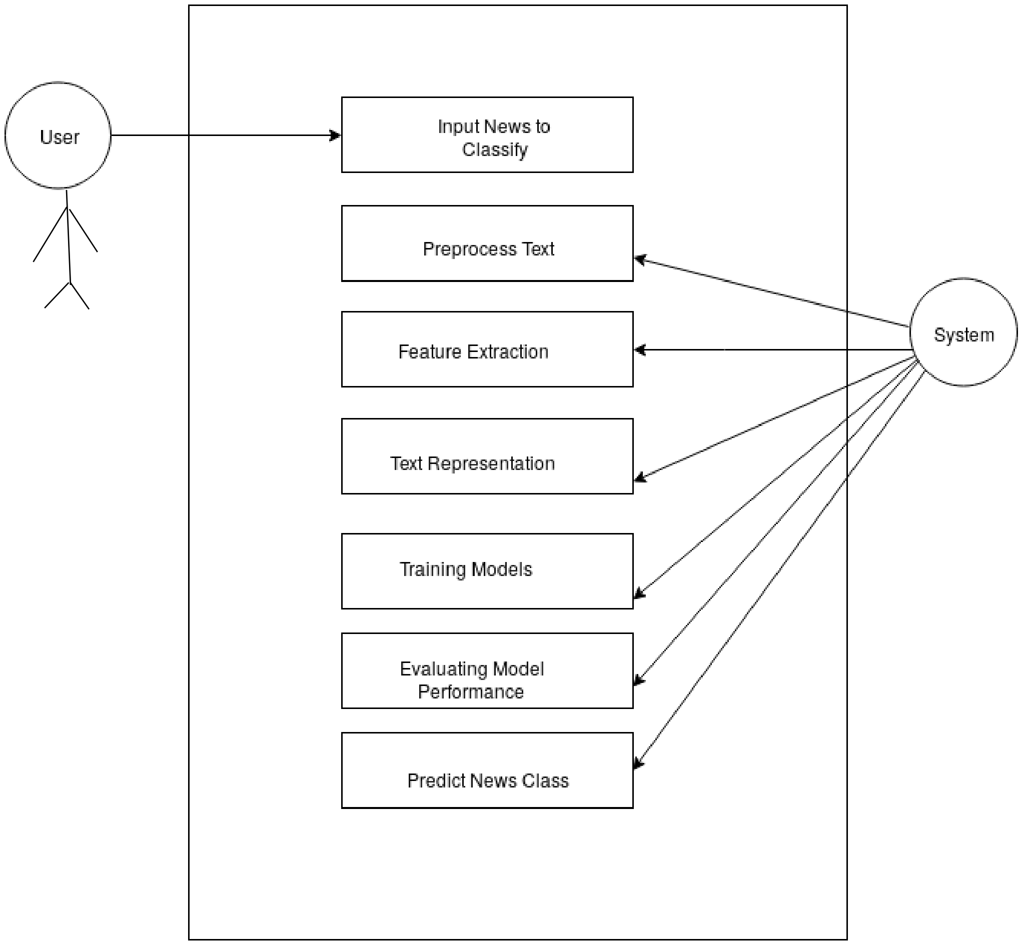


Figure 6 Use Case Diagram

3.3. Algorithms /Approaches Used for Classification

With the ever-increasing volume of text data from Internet, databases, and archives, text categorization or classification poses unique challenges due to the very high dimensionality of text data, sparsity, multi-class labels and unbalanced classes. In classification, the idea is to predict the target class by analysis the training dataset. This could be done by finding proper boundaries for each target class. In a general way of saying, Use the training dataset to get better boundary conditions which could be used to determine each target class. Once the boundary conditions determined, the next task is to predict the target class as we have said earlier. The whole process is known as classification**.** Therefore, this is what we are going to do in this project: Classifying Unclassified news articles into 12 pre-defined categories.

Many classification approaches have been developed for categorizing text documents, most of the researches in text categorization come from the machine learning and information retrieval communities such as decision trees, Random Forests, Naïve-Bayes (NB), Support Vector Machines (SVM), k-Nearest Neighbor (kNN), Gradient Boosting, Neural Network (NNet) and etc. We use the following algorithms in this project:

Support Vector Machine (SVM):

It is a machine learning approach that uses a linear classifier to classify data into two categories. The classifier is non-probabilistic. Support vector machine (SVM) is a supervised machine learning method capable of deciphering subtle patterns in noisy and complex datasets. Support vector machine technique creates a hyper plane in boundless dimensional space, which is classification and regression. A separation is accomplished by the hyper plane that has the biggest classification for the purpose of nearest training data point of any class. This nearest data point is known as functional margin. The generalization error of the SVM classifier depends on the size of the functional margin. The SVM training algorithm builds a model on the basis of functional margin; the category of functional margin makes a non-probabilistic binary linear classifier. This technique is a supervised learning model used for linear and non-linear classification. The non-linear classification is performed using kernel-based function for mapping input into high dimensional feature space.

Let the data set D be given as (X_1, y_1), (X_2, y_2) …. ($X_{|D|}$, $Y_{|D|}$), where X_i is the set of training tuples with associated class labels Y_i. Each y_i can take one of two values, either +1 or -1, A separating hyper-plane can be written as:

$$W.X+b=0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{ equation(1)}$$

W=weight vector i.e. W= {w_1, w_2, …. wn} where, n is the number of attributes and B is scalar, often referred as a bias and X=values of attributes [16]

SVM Classifiers attempt to partition the dataspace with the use of linear or non-linear delineations between the different classes. The key in such classifiers is to determine the optimal boundaries between the different classes and use them for the purposes of classification

KNN algorithm:

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. The algorithm is simple and easy to implement. There’s no need to build a model, tune several parameters, or make additional assumptions. The algorithm is versatile. It can be used for classification, regression, and search. The algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase. Moreover, there are faster algorithms that can produce more accurate classification and regression results. However, provided you have sufficient computing resources to speedily handle the data you are using to make predictions, KNN can still be useful in solving problems that have solutions that depend on identifying similar objects. An example of this is using the KNN algorithm in recommender systems, an application of KNN-search. The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It’s easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows. KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression). In the case of classification and regression, we saw that choosing the right K for our data is done by trying several Ks and picking the one that works best.

If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

Naïve Bayes Algorithm:

It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability. Naive Bayes model is easy to build and particularly useful for very large data sets. These are probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. It is and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. Bayes' theorem is stated mathematically as the following equation:

$$P(A/B) = \frac{P(B/A).P(A)}{P(B)} \text{.................equation (2)}$$

- Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as evidence.
- P(A) is the priori of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance (here, it is event B).
- P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.

The Naïve Bayes model is a family of classification models that makes what is called a naive assumption. The naive assumption is that attributes(words) are independent of each other. What this means is that the order of the attributes(words) does not matter. The Naïve Bayes model we used is in our experiments is commonly called a Multinomial Naïve Bayes model. This model takes into account the number of occurrences of an attribute in a document.

In Bayesian classifiers (also called generative classifiers), we attempt to build a probabilistic classifier based on modeling the underlying word features in different classes. The idea is then to classify text based on the posterior probability of the documents belonging to the different classes on the basis of the word presence in the documents.

Random Forest Algorithm:

Random forest algorithm is a supervised classification algorithm. Due to its algorithmic simplicity and prominent classification performance for high dimensional data, random forest has become a promising method for text categorization. Random forest is a popular classification method which is an ensemble of a set of classification trees. As the name suggest, this algorithm creates the forest with a number of trees. In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.

One of the most popular forest construction procedures, is to randomly select a subspace of features at each node to grow branches of a decision trees, then to use bagging method to generate training data subsets for building individual trees, finally to combine all individual trees to form random forests model. Text data has many terms or features which are uninformative to a specific topic (i.e., a class). During this forest building process, topic-related or informative features would have the large chance to be missed, if we randomly select a small subspace from high dimensional text data. As a result, weak trees will be created from these subspaces, the average discriminative strength of those trees in reduced and the error bound of the random forest is enlarged. Therefore, when a large proportion of such “weak” trees are generated in a random forest, the forest has a large likelihood to make a wrong decision which mainly results from those “weak” trees’ classification power.

Gradient Boosting Algorithm:

Boosting is a method of converting weak learners into strong learners. that work on the principle of boosting weak learners iteratively by shifting focus towards problematic observations that were difficult to predict in previous iterations and performing an ensemble of weak learners, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, but it generalizes them by allowing optimization of an arbitrary differentiable loss function. In boosting, each new tree is a fit on a modified version of the original data set. The gradient boosting algorithm (gbm) can be most easily explained by first introducing the AdaBoost Algorithm. The AdaBoost Algorithm begins by training a decision tree in which each observation is assigned an equal weight. After evaluating the first tree, we increase the weights of those observations that are difficult to classify and lower the weights for those that are easy to classify. The second tree is therefore grown on this weighted data.

Here, the idea is to improve upon the predictions of the first tree. Our new model is therefore *Tree 1 + Tree 2*. We then compute the classification error from this new 2-tree ensemble model and grow a third tree to predict the revised residuals. We repeat this process for a specified number of iterations. Subsequent trees help us to classify observations that are not well classified by the previous trees. Predictions of the final ensemble model is therefore the weighted sum of the predictions made by the previous tree models. Gradient Boosting trains many models in a gradual, additive and sequential manner.

The major difference between AdaBoost and Gradient Boosting Algorithm is how the two algorithms identify the shortcomings of weak learners (e.g. decision trees). While the gradient boosting performs the same by using gradients in the loss function as:

$$y= ax +b+ e\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{equation (3)}$$

where, 'e' is the error term.

The loss function is a measure indicating how good are model's coefficients are at fitting the underlying data. A logical understanding of loss function would depend on what we are trying to optimize. For example, if we are trying to predict the sales prices by using a regression, then the loss function would be based off the error between true and predicted house prices. Similarly, if our goal is to classify credit defaults, then the loss function would be a measure of how good our predictive model is at classifying bad loans. One of the biggest motivations of using gradient boosting is that it allows one to optimize a user specified cost function, instead of a loss function that usually offers less control and does not essentially correspond with real world applications.

Logistic Regression Algorithm:

Logistic regression is named for the function used at the core of the method, the logistic function. The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits. When it comes to classification, we are determining the probability of an observation to be part of a certain class or not. Therefore, we wish to express the probability with a value between 0 and 1. A probability close to 1 means the observation is very likely to be part of that category. In order to generate values between 0 and 1, we express the probability using this equation of logistic function. Making predictions with a logistic regression model is as simple as plugging in numbers into the logistic regression equation and calculating a result. The logistic function, can write our class probability like below.

$$P\left(\frac{Y}{X}\right) = \frac{1}{1+e^{-f(x)}} \text{.............................. equation (4)}$$

where *f(x)* is a function consisting our features (x_j) and their corresponding weights/coefficients (β_j) in a linear form shown below.

$$f(x)=x_0+x_1\beta_1+\ldots + x_k\beta_k+ \varepsilon \text{ .... equation (5)}$$

where *x, β, f(x)* ϵR^k and *ε* is representing the '*random error process — noise'* inevitably happening in the data generating process.

Classification algorithms do what the name suggests i.e. they train models to predict what class some object belongs to. A very common application News Classification. Logistic Regression is an algorithm that is relatively simple and powerful for deciding between two classes, i.e. it's a binary classifier. It basically gives a function that is a boundary between two different classes. It can be extended to handle more than two classes by a method referred to as "one-vs-all" (multinomial logistic regression or softmax regression) which is really a collection of binary classifiers that just picks out the most likely class by looking at each class individually verses everything else and then picks the class that has the highest probability.

It has observed that even for a specified classification method, classification performances of the classifiers based on different training text corpuses are different; and in some cases, such differences are quite substantial. This literature study observation implies that:

a) *classifier performance is relevant to its training data in some degree,* and

b) *good or high-quality training data may derive classifiers of good performance*.

Unfortunately, up to now little research work in the literature has been seen on how to exploit the problems in existing approaches to improve classifier's performance.

3.4. Software Development Life Cycle: Incremental

The framework we will be using for developing this project is Incremental model. Incremental Model is a process of software development where requirements are broken down into multiple standalone modules of software development cycle. This model combines linear sequential model with the iterative prototype model. They all fundamentally incorporate iteration and the continuous feedback that it provides to successively refine and deliver a software system. New functionalities will be added as each increment is developed. The major phases of the linear sequential model are: Analysis, Design, Coding and Testing. The software repeatedly passes through these phases in iteration and an increment is delivered with progressive changes.

In general, an SDLC methodology follows these following steps:

1. Analysis: The existing systems related to NLP domain, are evaluated. We consulted with our teachers, took suggestions from seniors and performed series of intense internet researches on this domain to come up with this project idea, analyze its feasibility, scope and significances. The new system requirements are defined. In particular, the deficiencies in the existing system must be addressed with specific proposals for improvement. Other factors defined include needed features, functions and capabilities. Since the project is a problem of multiclass classification, 12 categories of news are considered, which can cover most of the subjective domains of news. Then the requirement is to collect news from each category, in order to train the models. Different Nepali News portals were scraped to gather the required categorical news.

2.Design: After the analysis of the data collected, model parameters obtained at each iteration and the performance accuracy of each increment, the system is designed. Plans are laid out concerning the hardware(pc), operating systems(windows/ubuntu), programming (python) and communications issues. The GUI design was changed after the first iteration to obtain more interactive and easier to use interface.

3.Code: The new system is developed. The models are trained in Jupyter notebook. As the trained models are to be implemented, a user-friendly GUI is then developed using Tkinter library of Python. The new components and programs are obtained and installed. Users of the system must be trained in its use.

4.Test: The system is incorporated in a production environment. All aspects of performance must be tested. If necessary, adjustments must be made at this stage. This step involves changing and updating the system once it is in place. Updates in the code are, in some way to better fit the needs of the end-users continuously. The performance values of each algorithmic models are compared and analyzed to determine best classifier. The application was tested with untrained set of articles.

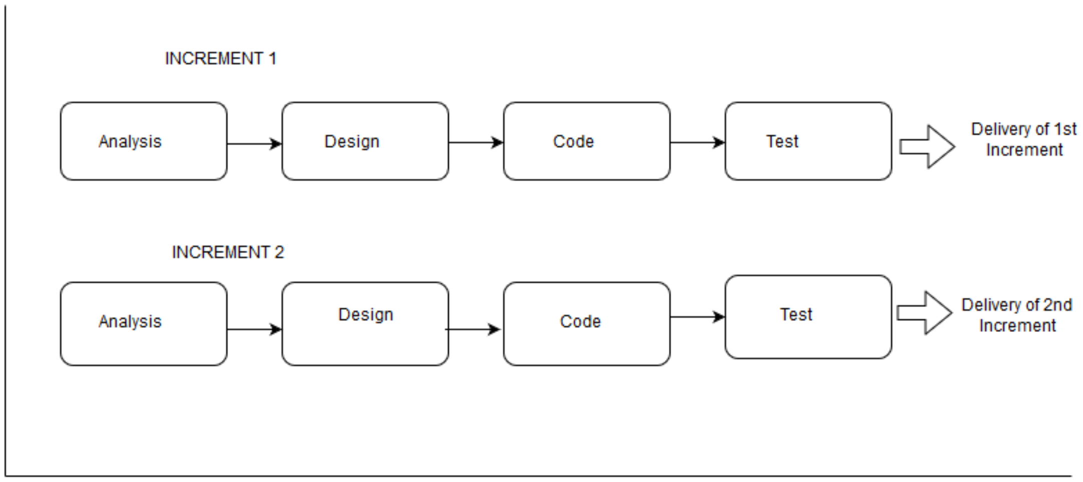


Figure 7: Incremental Software Development Life Cycle

3.5. Technologies Used

- Programming Languages: Python
- Libraries:
  - ➢ NLTK- Natural Language Tool Kit is a popular library for NLP tasks, containing required solutions for different NLP problems
  - ➢ Re- It is python library for Regular Expression Handling, used for pattern matching
  - ➢ Tkinter- It is Graphical User Interface Application Library for Python
  - ➢ Scikit-learn- It is python library for Scientific Processing used for Language Features [17]
  - ➢ Pandas- Library for Data Frames and arranging datasets in tabular formats
  - ➢ Numpy- Python library for numerical calculation
  - ➢ BeautifulSoup- Web Scrapping [18] Library
  - ➢ Requests- HTML parser library from web pages

3.6. Tools Used

The tools used for documentation, designing and developing the application, testing and deployment are listed in the table below:

TOOLS	PURPOSE
MS Excel/ Notepad	Data Collection, Dataset Formation
Jupyter Notebook, Spyder	IDE (Interactive Development Environment) for data analysis and model training
Python IDLE	Coding for GUI

Table 1: Tools used

3.7. Performance Analysis Methodology and Validation Scheme

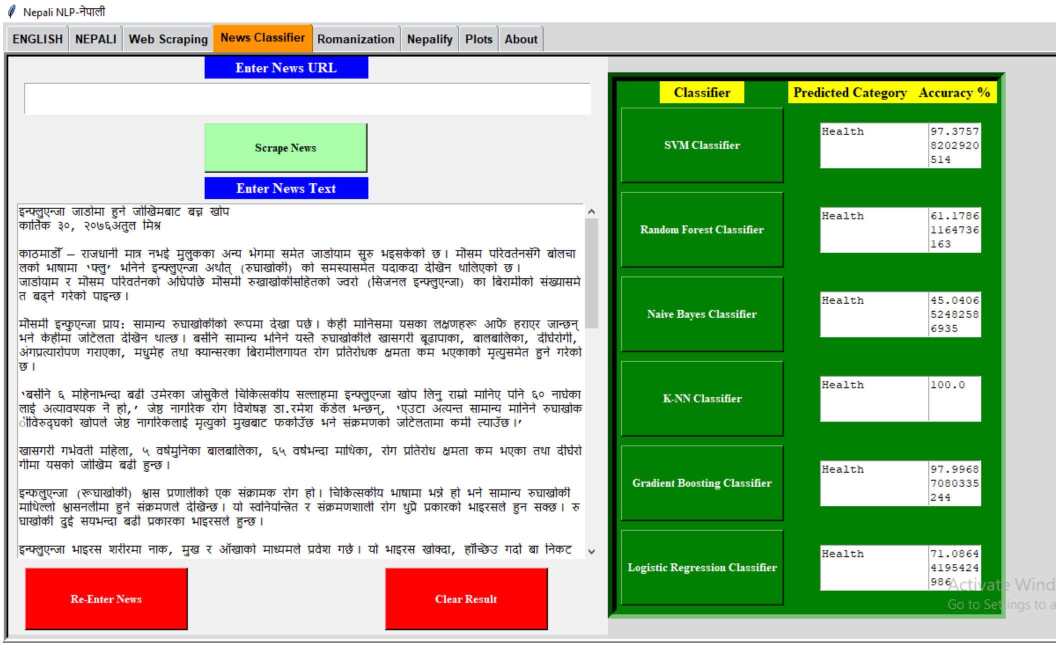
The evaluation of the efficiency of our Classification Models are done using cross-validation and resampling techniques that are commonly used to derive point estimates of the performances which are compared to identify methods with good properties. Predictive accuracy of the models are evaluated using the performance metrices from the classification report of scikit-learn.

At the completion of project, the evaluation is done based on performance metrics of the deliverable obtained.

4. RESULT

As the outcome of this project, we obtain an application that can perform with meeting the proposed objectives of our project, that is a desktop Application. This system has been capable of performing Nepali News Classification using six different Algorithms and analyzing and comparing the performance of them. Along with some other NLP related tasks like POS tagging, tokenization, Nepali transliteration and romanization etc.

Some of the key results of this projects are illustrated below:



इन्फ्लुएन्जा जाडोमा हुने जोखिमबाट बच्न खोप
कार्तिक ३०, २०७६अतुल मिश्र

काठमाडौं – राजधानी मात्र नभई मुलुकका अन्य भेगमा समेत जाडोयाम सुरु भइसकेको छ। मौसम परिवर्तनसँगै बोलचा
लको भाषामा ‘फ्लु’ भनिने इन्फ्लुएन्जा अर्थात् (रुघाखोकी) को समस्यासमेत यदाकदा देखिन थालिएको छ।
जाडोयाम र मौसम परिवर्तनको आघेपाछे मौसमी रुखाखोकीसहितको ज्वरो (सिजनल इन्फ्लुएन्जा) का बिरामीको संख्यासमे
त बढ्ने गरेको पाइन्छ।

मौसमी इन्फुएन्जा प्रायः सामान्य रुघाखोकीको रूपमा देखा पर्छ। केही मानिसमा यसका लक्षणहरू आफै हराएर जान्छन्
भने केहीमा जटिलता देखिन थाल्छ। बसौंने सामान्य भनिने यस्तै रुघाखोकीले खासगरी बूढापाका, बालबालिका, दीर्घरोगी,
अंगप्रत्यारोपण गराएका, मधुमेह तथा क्यान्सरका बिरामीलगायत रोग प्रतिरोधक क्षमता कम भएकाको मृत्युसमेत हुने गरेको
छ।

‘बसौंने ६ महिनाभन्दा बढी उमेरका जोसुकैले चिकित्सकीय सल्लाहमा इन्फ्लुएन्जा खोप लिनु राम्रो मानिए पनि ६० नाघेका
लाई अत्यावश्यक नै हो,’ जेष्ठ नागरिक रोग विशेषज्ञ डा.रमेश कँडेल भन्छन्, ‘एउटा अत्यन्त सामान्य मानिने रुघाखोक
ीविरुद्धको खोपले जेष्ठ नागरिकलाई मृत्युको मुखबाट फर्काउँछ भने संक्रमणको जटिलतामा कमी ल्याउँछ।’

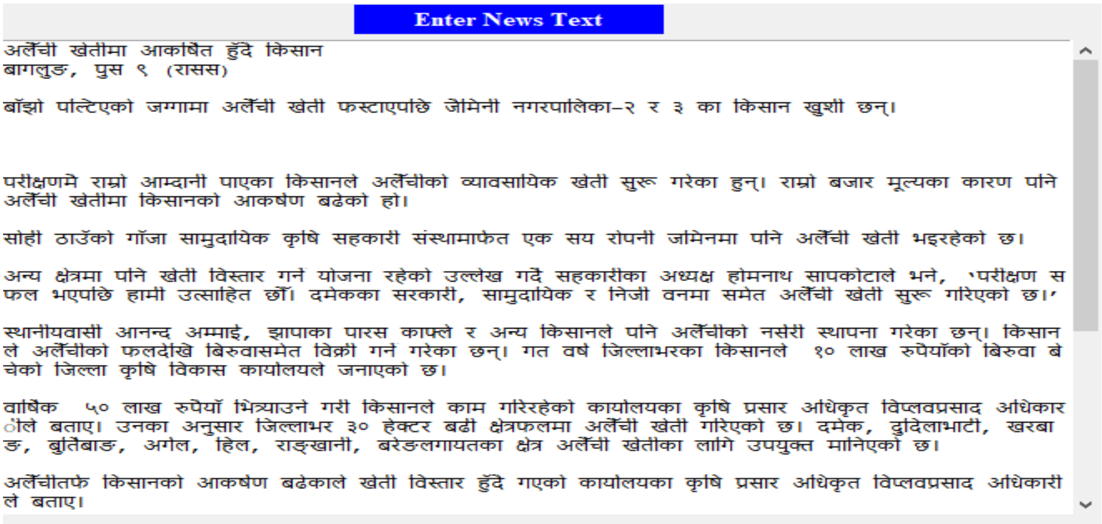
खासगरी गर्भवती महिला, ५ वर्षमुनिका बालबालिका, ६५ वर्षभन्दा माथिका, रोग प्रतिरोध क्षमता कम भएका तथा दीर्घरो
गीमा यसको जोखिम बढी हुन्छ।

इन्फलुएन्जा (रूघाखोकी) श्वास प्रणालीको एक संक्रामक रोग हो। चिकित्सकीय भाषामा भन्ने हो भने सामान्य रुघाखोकी
माथिल्लो श्वासनलीमा हुने संक्रमणले देखिन्छ। यो स्वनियन्त्रित र संक्रमणशाली रोग धुप्रै प्रकारको भाइरसले हुन सक्छ। रु
घाखोकी दुई सयभन्दा बढी प्रकारका भाइरसले हुन्छ।

इन्फ्लुएन्जा भाइरस शरीरमा नाक, मुख र आँखाको माध्यमले प्रवेश गर्छ। यो भाइरस खोक्दा, हाँच्छिउ गर्दा वा निकट

Classifier	Predicted Category	Accuracy %
SVM Classifier	Health	97.3757 8202920 514
Random Forest Classifier	Health	61.1786 1164736 163
Naive Bayes Classifier	Health	45.0406 5248258 6935
K-NN Classifier	Health	100.0
Gradient Boosting Classifier	Health	97.9968 7080335 244
Logistic Regression Classifier	Health	71.0864 4195424 986

Figure 8: Application window with classification interfaces



अलैंची खेतीमा आकर्षित हुँदै किसान
बागलुङ, पुस ९ (रासस)

बाँझो पल्टिएको जग्गामा अलैंची खेती फस्टाएपछि जैमिनी नगरपालिका–२ र ३ का किसान खुशी छन्।

परीक्षणमै राम्रो आम्दानी पाएका किसानले अलैंचीको व्यावसायिक खेती सुरू गरेका हुन्। राम्रो बजार मूल्यका कारण पनि
अलैंची खेतीमा किसानको आकर्षण बढेको हो।

सोही ठाउँको गाँजा सामुदायिक कृषि सहकारी संस्थामार्फत एक सय रोपनी जमिनमा पनि अलैंची खेती भइरहेको छ।

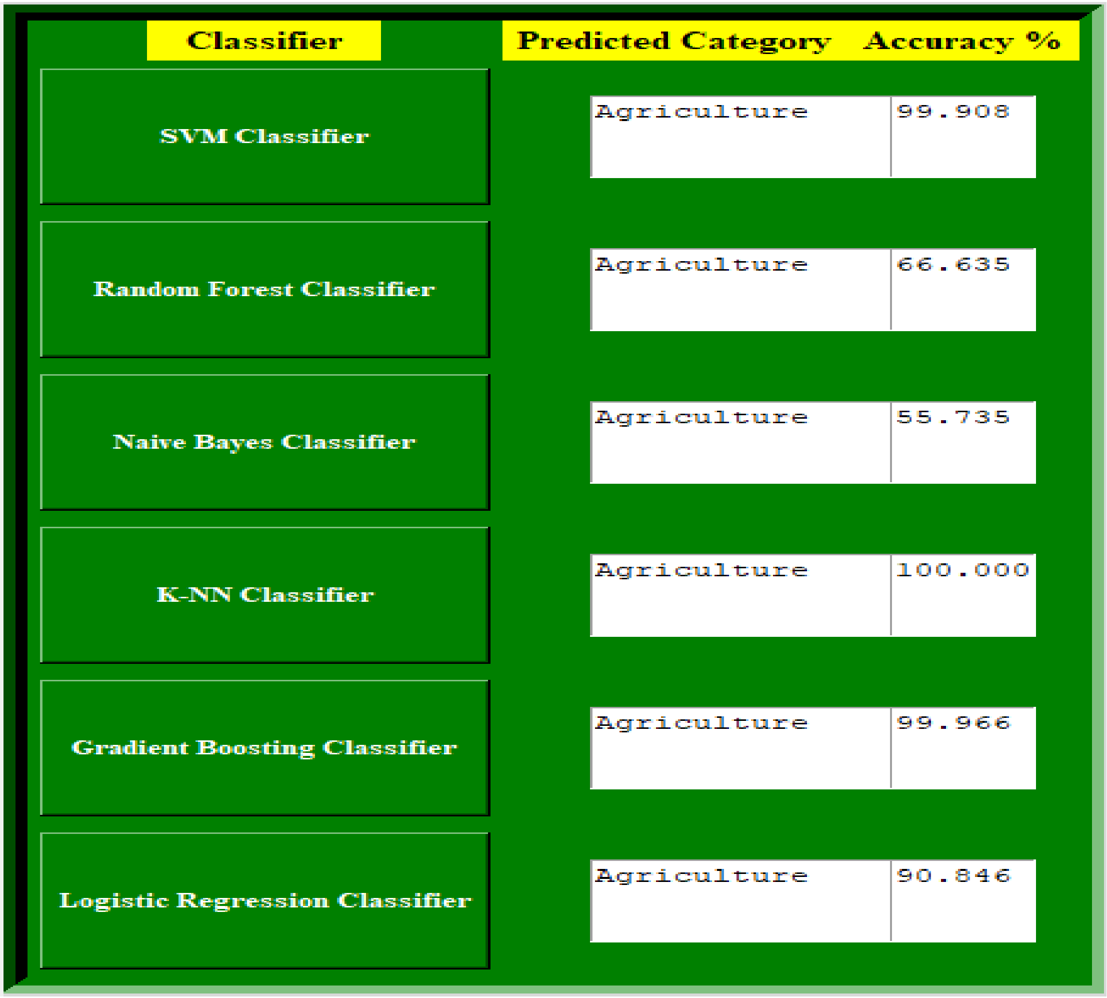
अन्य क्षेत्रमा पनि खेती विस्तार गर्ने योजना रहेको उल्लेख गर्दै सहकारीका अध्यक्ष होमनाथ सापकोटाले भने, ‘परीक्षण स
फल भएपछि हामी उत्साहित छौँ। दमेकका सरकारी, सामुदायिक र निजी वनमा समेत अलैंची खेती सुरू गरिएको छ।’

स्थानीयवासी आनन्द अम्माइ, झापाका पारस काफ्ले र अन्य किसानले पनि अलैंचीको नर्सरी स्थापना गरेका छन्। किसान
ले अलैंचीको फलदेखि बिरुवासमेत बिक्री गर्ने गरेका छन्। गत वर्ष जिल्लाभरका किसानले १० लाख रुपैयाँको बिरुवा बे
चेको जिल्ला कृषि विकास कार्यालयले जनाएको छ।

वार्षिक ५० लाख रुपैयाँ भित्र्याउने गरी किसानले काम गरिरहेको कार्यालयका कृषि प्रसार अधिकृत विप्लवप्रसाद अधिकार
ीले बताए। उनका अनुसार जिल्लाभर ३० हेक्टर बढी क्षेत्रफलमा अलैंची खेती गरिएको छ। दमेक, दुदिलाभाटी, खरबा
ङ, बुर्तिबाङ, अगेल, हिल, राङ्खानी, बरेङलगायतका क्षेत्र अलैंची खेतीका लागि उपयुक्त मानिएको छ।

अलैंचीतर्फ किसानको आकर्षण बढेकाले खेती विस्तार हुँदै गएको कार्यालयका कृषि प्रसार अधिकृत विप्लवप्रसाद अधिकारी
ले बताए।

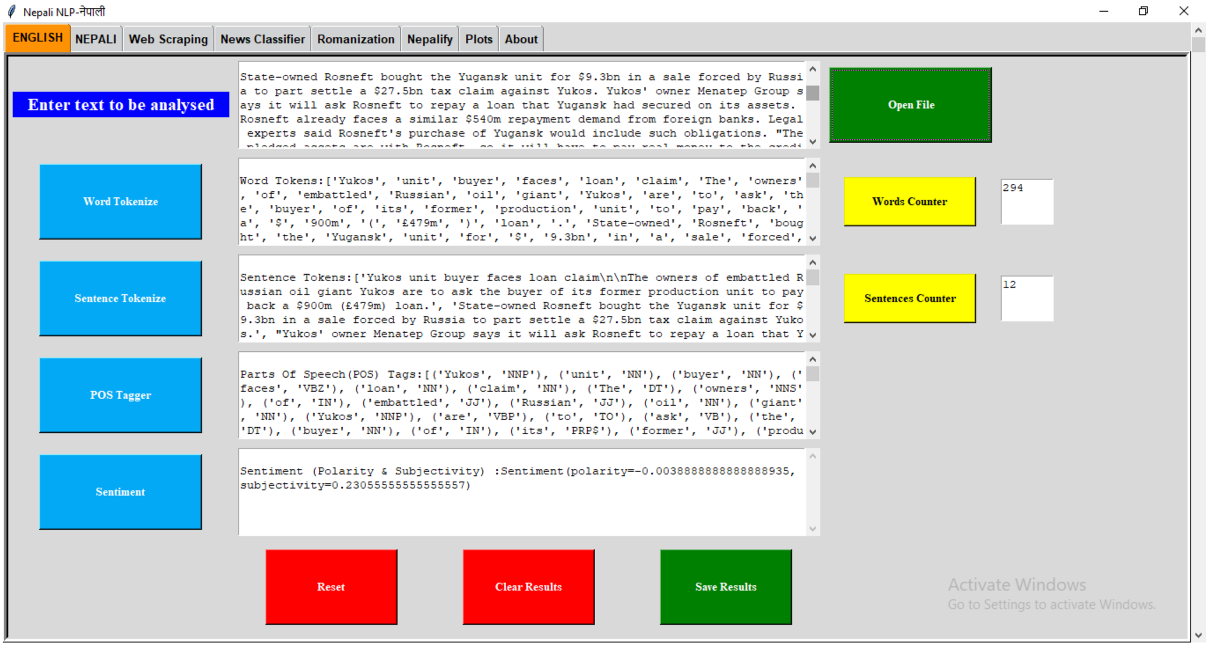
Figure 9: The news input interface with a sample test news inserted



Classifier	Predicted Category	Accuracy %
SVM Classifier	Agriculture	99.908
Random Forest Classifier	Agriculture	66.635
Naive Bayes Classifier	Agriculture	55.735
K-NN Classifier	Agriculture	100.000
Gradient Boosting Classifier	Agriculture	99.966
Logistic Regression Classifier	Agriculture	90.846

Figure 10: The Category prediction Interface with Result

Figure 10 shows the interface for displaying the predicted category and its accuracy of prediction for different classifier models. The News inserted in figure, is Actually of category Agriculture. All the classifiers have predicted this correctly with different score of accuracy.



```
State-owned Rosneft bought the Yugansk unit for $9.3bn in a sale forced by Russi
a to part settle a $27.5bn tax claim against Yukos. Yukos' owner Menatep Group s
ays it will ask Rosneft to repay a loan that Yugansk had secured on its assets.
Rosneft already faces a similar $540m repayment demand from foreign banks. Legal
 experts said Rosneft's purchase of Yugansk would include such obligations. "The
```

Words Counter: 294

Sentences Counter: 12

```
Word Tokens:['Yukos', 'unit', 'buyer', 'faces', 'loan', 'claim', 'The', 'owners'
, 'of', 'embattled', 'Russian', 'oil', 'giant', 'Yukos', 'are', 'to', 'ask', 'th
e', 'buyer', 'of', 'its', 'former', 'production', 'unit', 'to', 'pay', 'back', '
a', '$', '900m', '(', '£479m', ')', 'loan', '.', 'State-owned', 'Rosneft', 'boug
ht', 'the', 'Yugansk', 'unit', 'for', '$', '9.3bn', 'in', 'a', 'sale', 'forced',
```

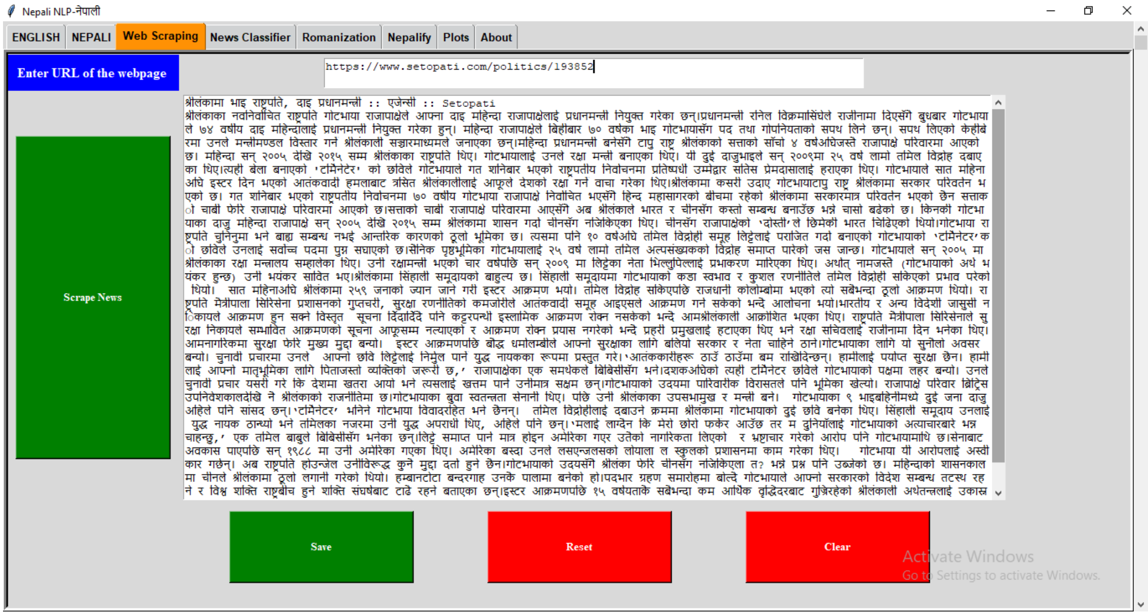
```
Sentence Tokens:['Yukos unit buyer faces loan claim\n\nThe owners of embattled R
ussian oil giant Yukos are to ask the buyer of its former production unit to pay
 back a $900m (£479m) loan.', 'State-owned Rosneft bought the Yugansk unit for $
9.3bn in a sale forced by Russia to part settle a $27.5bn tax claim against Yuko
s.', "Yukos' owner Menatep Group says it will ask Rosneft to repay a loan that Y
```

```
Parts Of Speech(POS) Tags:[('Yukos', 'NNP'), ('unit', 'NN'), ('buyer', 'NN'), ('
faces', 'VBZ'), ('loan', 'NN'), ('claim', 'NN'), ('The', 'DT'), ('owners', 'NNS'
), ('of', 'IN'), ('embattled', 'JJ'), ('Russian', 'JJ'), ('oil', 'NN'), ('giant'
, 'NN'), ('Yukos', 'NNP'), ('are', 'VBP'), ('to', 'TO'), ('ask', 'VB'), ('the',
'DT'), ('buyer', 'NN'), ('of', 'IN'), ('its', 'PRP$'), ('former', 'JJ'), ('produ
```

```
Sentiment (Polarity & Subjectivity) :Sentiment(polarity=-0.003888888888888935,
subjectivity=0.23055555555555557)
```

Figure 11: Interface with Some Basic NLP Tasks Results

Figure 11 shows the additional task that can be performed with this application. This shows the interface which can perform basic NLP tasks like word tokenization, sentences tokenization, POS tagging, and Sentiment Analysis using libraries like nltk, textblob and spacy.



```
https://www.setopati.com/politics/193852
```

```
श्रीलंकामा भाइ राष्ट्रपति, दाइ प्रधानमन्त्री :: एजेन्सी :: Setopati
श्रीलंकाका नवनिर्वाचित राष्ट्रपति गोटभाया राजापाक्षेले आफ्ना दाइ महिन्दा राजापाक्षेलाइ प्रधानमन्त्री नियुक्त गरेका छन्।प्रधानमन्त्री रनिल विक्रमासिंघेले राजीनामा दिएसँगै बुधबार गोटभाया
ले ७४ वर्षीय दाइ महिन्दालाई प्रधानमन्त्री नियुक्त गरेका हुन्। महिन्दा राजापाक्षेले बिहीबार ७० वर्षका भाइ गोटभायासँग पद तथा गोपनियताको सपथ लिने छन्। सपथ लिएको केहीबे
रमा उनले मन्त्रीमण्डल विस्तार गर्ने श्रीलंकाली सञ्चारमाध्यमले जनाएका छन्।महिन्दा प्रधानमन्त्री बनेसँगै टापु राष्ट्र श्रीलंकाको सत्ताको साँचो ४ वर्षअघिजस्तै राजापाक्षे परिवारमा आएको
छ। महिन्दा सन् २००५ देखि २०१५ सम्म श्रीलंकाका राष्ट्रपति थिए। गोटभायालाई उनले रक्षा मन्त्री बनाएका थिए। यी दुई दाजुभाइले सन् २००९मा २५ वर्ष लामो तामिल विद्रोह दबाए
का थिए।त्यही बेला बनाएको 'टर्मिनेटर' को छविले गोटभायाले गत शनिबार भएको राष्ट्रपतीय निर्वाचनमा प्रतिष्पर्धी उम्मेदवार सजित प्रेमदासालाई हराएका थिए। गोटभायाले सात महिना
अघि इस्टर दिन भएको आतंकवादी हमलाबाट त्रसित श्रीलंकालीलाई आफूले देशको रक्षा गर्ने वाचा गरेका थिए।श्रीलंकामा कसरी उदाए गोटभायाटापु राष्ट्र श्रीलंकामा सरकार परिवर्तन भ
एको छ। गत शनिबार भएको राष्ट्रपतीय निर्वाचनमा ७० वर्षीय गोटभाया राजापाक्षे निर्वाचित भएसँगै हिन्द महासागरको बीचमा रहेको श्रीलंकामा सरकारमात्र परिवर्तन भएको छैन सत्ताक
ो चाबी फेरि राजापाक्षे परिवारमा आएको छ।सत्ताको चाबी राजापाक्षे परिवारमा आएसँगै अब श्रीलंकाले भारत र चीनसँग कस्तो सम्बन्ध बनाउँछ भन्ने चासो बढेको छ। किनकी गोटभा
याका दाजु महिन्दा राजापाक्षे सन् २००५ देखि २०१५ सम्म श्रीलंकामा शासन गर्दा चीनसँग नजिकिएका थिए। चीनसँग राजापाक्षेको 'दोस्ती' ले छिमेकी भारत चिढिएको थियो।गोटभाया रा
ष्ट्रपति चुनिनुमा भने बाह्य सम्बन्ध नभई आन्तरिक कारणको ठूलो भूमिका छ। त्यसमा पनि १० वर्षअघि तामिल विद्रोही समूह लिट्टेलाई पराजित गर्दा बनाएको गोटभायाको 'टर्मिनेटर' क
ो छविले उनलाई सर्वोच्च पदमा पुग्न सघाएको छ।सैनिक पृष्ठभूमिका गोटभायालाई २५ वर्ष लामो तामिल अल्पसंख्यकको विद्रोह समाप्त पारेको जस जान्छ। गोटभायाले सन् २००५ मा
श्रीलंकाका रक्षा मन्त्रालय सम्हालेका थिए। उनी रक्षामन्त्री भएको चार वर्षपछि सन् २००९ मा लिट्टेका नेता भिल्लुपिल्लाई प्रभाकरण मारिएका थिए। अर्थात् नामजस्तै (गोटभायाको अर्थ भ
यंकर हुन्छ) उनी भयंकर साबित भए।श्रीलंकामा सिंहाली समुदायको बाहुल्य छ। सिंहाली समुदायमा गोटभायाको कडा स्वभाव र कुशल रणनीतिले तामिल विद्रोही सकिएको प्रभाव परेको
 थियो।  सात महिनाअघि श्रीलंकामा २५९ जनाको ज्यान जाने गरी इस्टर आक्रमण भयो। तामिल विद्रोह सकिएपछि राजधानी कोलोम्बोमा भएको त्यो सबैभन्दा ठूलो आक्रमण थियो। रा
ष्ट्रपति मैत्रीपाला सिरिसेना प्रशासनको गुप्तचरी, सुरक्षा रणनीतिको कमजोरीले आतंकवादी समूह आइएसले आक्रमण गर्न सकेको भन्दै आलोचना भयो।भारतीय र अन्य विदेशी जासुसी न
िकायले आक्रमण हुन सक्ने विस्तृत  सूचना दिँदादिँदै पनि कट्टरपन्थी इस्लामिक आक्रमण रोक्न नसकेको भन्दै आमश्रीलंकाली आक्रोशित भएका थिए। राष्ट्रपति मैत्रीपाला सिरिसेनाले सु
रक्षा निकायले सम्भावित आक्रमणको सूचना आफूसम्म नल्याएको र आक्रमण रोक्न प्रयास नगरेको भन्दै प्रहरी प्रमुखलाई हटाएका थिए भने रक्षा सचिवलाई राजीनामा दिन भनेका थिए।
आमनागरिकमा सुरक्षा फेरि मुख्य मुद्दा बन्यो।  इस्टर आक्रमणपछि बौद्ध धर्मावलम्बीले आफ्नो सुरक्षाका लागि बलियो सरकार र नेता चाहिने ठाने।गोटभायाका लागि यो सुनौलो अवसर
बन्यो। चुनावी प्रचारमा उनले  आफ्नो छवि लिट्टेलाई निर्मूल पार्ने युद्ध नायकका रूपमा प्रस्तुत गरे।'आतंककारीहरू ठाउँ ठाउँमा बम राखिदिन्छन्। हामीलाई पर्याप्त सुरक्षा छैन। हामी
लाई आफ्नो मातृभूमिका लागि पिताजस्तो व्यक्तिको जरूरी छ,' राजापाक्षेका एक समर्थकले बिबिसीसँग भने।दशकअघिको त्यही टर्मिनेटर छविले गोटभायाको पक्षमा लहर बन्यो। उनले
चुनावी प्रचार यसरी गरे कि देशमा खतरा आयो भने त्यसलाई खत्तम पार्न उनीमात्र सक्षम छन्।गोटभायाको उदयमा पारिवारीक विरासतले पनि भूमिका खेल्यो। राजापाक्षे परिवार ब्रिटिस
उपनिवेशकालदेखि नै श्रीलंकाको राजनीतिमा छ।गोटभायाका बुवा स्वतन्त्रता सेनानी थिए। पछि उनी श्रीलंकाका उपसभामुख र मन्त्री बने।  गोटभायाका ९ भाइबहिनीमध्ये दुई जना दाजु
अहिले पनि सांसद छन्।'टर्मिनेटर' भनिने गोटभाया विवादरहित भने छैनन्।  तामिल विद्रोहीलाई दबाउने क्रममा श्रीलंकामा गोटभायाको दुई छवि बनेका थिए। सिंहाली समुदाय उनलाई
 युद्ध नायक ठान्थ्यो भने तामिलका नजरमा उनी युद्ध अपराधी थिए, अहिले पनि छन्।'मलाई लाग्दैन कि मेरो छोरो फर्केर आउँछ तर म दुनियाँलाई गोटभायाको अत्याचारबारे भन्न
चाहन्छु,' एक तामिल बाबुले बिबिसीसँग भनेका छन्।लिट्टे समाप्त पार्न मात्र होइन अमेरिका गएर उतैको नागरिकता लिएको  र भ्रष्टाचार गरेको आरोप पनि गोटभायामाथि छ।सेनाबाट
अवकास पाएपछि सन् १९८८ मा उनी अमेरिका गएका थिए। अमेरिका बस्दा उनले लसएन्जलसको लोयाला ल स्कुलको प्रशासनमा काम गरेका थिए।   गोटभाया यी आरोपलाई अस्वी
कार गर्छन्। अब राष्ट्रपति होउन्जेल उनीविरूद्ध कुनै मुद्दा दर्ता हुने छैन।गोटभायाको उदयसँगै श्रीलंका फेरि चीनसँग नजिकिएला त? भन्ने प्रश्न पनि उब्जेको छ। महिन्दाको शासनकाल
मा चीनले श्रीलंकामा ठूलो लगानी गरेको थियो। हम्बानटोटा बन्दरगाह उनकै पालामा बनेको हो।पदभार ग्रहण समारोहमा बोल्दै गोटभायाले आफ्नो सरकारको विदेश सम्बन्ध तटस्थ रह
ने र विश्व शक्ति राष्ट्रबीच हुने शक्ति संघर्षबाट टाढै रहने बताएका छन्।इस्टर आक्रमणपछि १५ वर्षयताकै सबैभन्दा कम आर्थिक वृद्धिदरबाट गुज्रिरहेको श्रीलंकाली अर्थतन्त्रलाई उकास्न
```

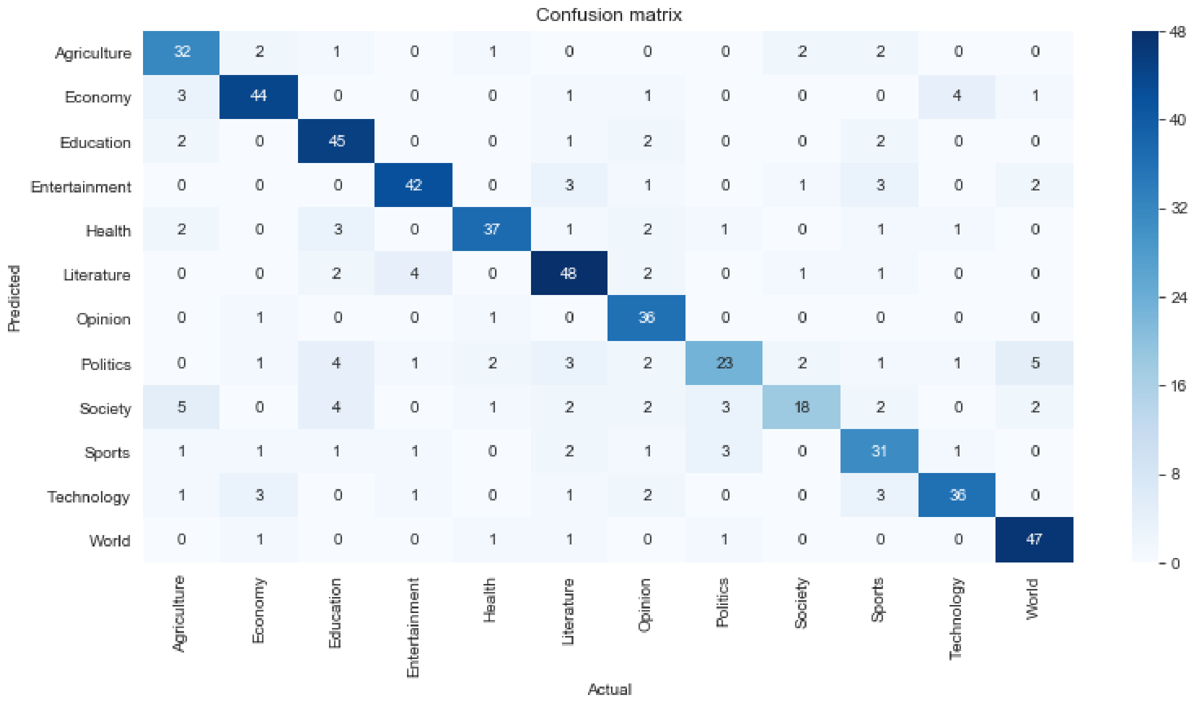
Figure 12: News text extraction using web scraping

Figure 12 the additional feature of this application, where news texts can be obtained and can be stored from the url provided. This can be applicable only for some Nepali news portals like www.setopati.com and www.ekantipur.com .

Confusion Matrix:

Confusion Matrix is the performance metric used for evaluating the models in the test datasets. In this project, for all the six classifier models, the confusion matrices are obtained as bellow, which shows the relations of predicted class label and actual class label, for some sample test data.
Taking random test data from the test dataset, and evaluating the prediction the results below are obtained, where total test samples, corrected predicted samples and incorrect samples are shown, for each category.

Gradient Boosting Classifier



	Agriculture	Economy	Education	Entertainment	Health	Literature	Opinion	Politics	Society	Sports	Technology	World
Agriculture	32	2	1	0	1	0	0	0	2	2	0	0
Economy	3	44	0	0	0	1	1	0	0	0	4	1
Education	2	0	45	0	0	1	2	0	0	2	0	0
Entertainment	0	0	0	42	0	3	1	0	1	3	0	2
Health	2	0	3	0	37	1	2	1	0	1	1	0
Literature	0	0	2	4	0	48	2	0	1	1	0	0
Opinion	0	1	0	0	1	0	36	0	0	0	0	0
Politics	0	1	4	1	2	3	2	23	2	1	1	5
Society	5	0	4	0	1	2	2	3	18	2	0	2
Sports	1	1	1	1	0	2	1	3	0	31	1	0
Technology	1	3	0	1	0	1	2	0	0	3	36	0
World	0	1	0	0	1	1	0	1	0	0	0	47

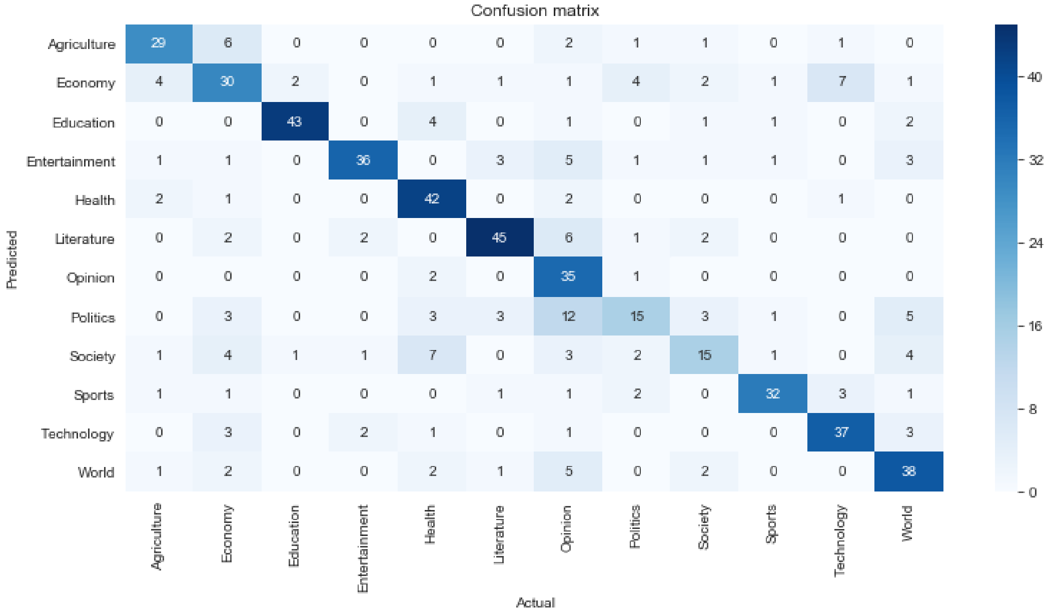
Figure 13: Confusion Matrix for Gradient Boosting Classifier

Categories	Test Samples	Predicted Correctly	Predicted Incorrectly
Agriculture	46	32	14
Education	60	45	15
Economy	53	44	9
Entertainment	49	42	7
Technology	43	36	7
World	57	47	10
Opinion	51	36	15
Politics	31	23	8
Society	24	18	6
Sports	46	31	15
Literature	63	48	15
Health	43	37	6

Table 2 Test Result for Gradient Boosting Classifier

Therefore, Accuracy= $\frac{CorrectlyPredictedSamples}{TotalTestSamples}$
=439/566=0.775%

K-NN Classifier



	Agriculture	Economy	Education	Entertainment	Health	Literature	Opinion	Politics	Society	Sports	Technology	World
Agriculture	29	6	0	0	0	0	2	1	1	0	1	0
Economy	4	30	2	0	1	1	1	4	2	1	7	1
Education	0	0	43	0	4	0	1	0	1	1	0	2
Entertainment	1	1	0	36	0	3	5	1	1	1	0	3
Health	2	1	0	0	42	0	2	0	0	0	1	0
Literature	0	2	0	2	0	45	6	1	2	0	0	0
Opinion	0	0	0	0	2	0	35	1	0	0	0	0
Politics	0	3	0	0	3	3	12	15	3	1	0	5
Society	1	4	1	1	7	0	3	2	15	1	0	4
Sports	1	1	0	0	0	1	1	2	0	32	3	1
Technology	0	3	0	2	1	0	1	0	0	0	37	3
World	1	2	0	0	2	1	5	0	2	0	0	38

Figure 14: Confusion Matrix for knn Classifier

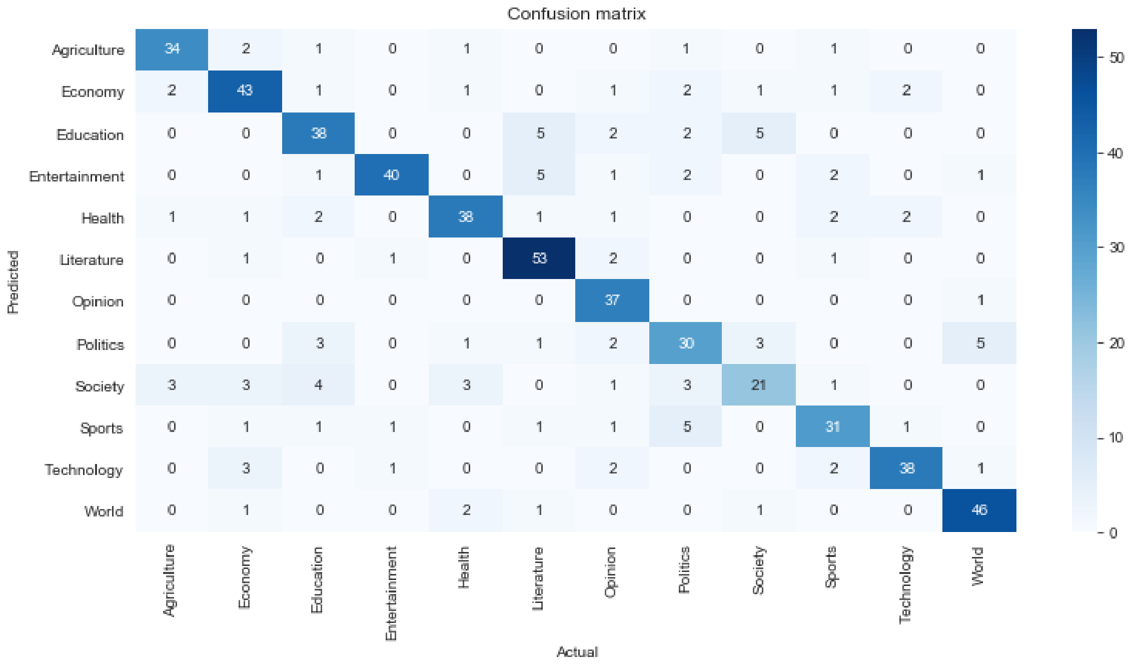
Categories	Test Samples	Predicted Correctly	Predicted Incorrectly
Agriculture	39	29	10
Education	46	43	3
Economy	53	30	23
Entertainment	41	36	5
Technology	49	37	12
World	57	38	19
Opinion	74	35	39
Politics	27	15	12
Society	27	15	12
Sports	37	32	5
Literature	54	45	9
Health	62	42	20

Table 3: Test Result for k-NN Classifier

Therefore, Accuracy= $\frac{CorrectlyPredictedSamples}{TotalTestSamples}$

=397/566=0.701%

Logistic Regression Classifier



	Agriculture	Economy	Education	Entertainment	Health	Literature	Opinion	Politics	Society	Sports	Technology	World
Agriculture	34	2	1	0	1	0	0	1	0	1	0	0
Economy	2	43	1	0	1	0	1	2	1	1	2	0
Education	0	0	38	0	0	5	2	2	5	0	0	0
Entertainment	0	0	1	40	0	5	1	2	0	2	0	1
Health	1	1	2	0	38	1	1	0	0	2	2	0
Literature	0	1	0	1	0	53	2	0	0	1	0	0
Opinion	0	0	0	0	0	0	37	0	0	0	0	1
Politics	0	0	3	0	1	1	2	30	3	0	0	5
Society	3	3	4	0	3	0	1	3	21	1	0	0
Sports	0	1	1	1	0	1	1	5	0	31	1	0
Technology	0	3	0	1	0	0	2	0	0	2	38	1
World	0	1	0	0	2	1	0	0	1	0	0	46

Figure 15: Confusion Matrix for Logistic Regression Classifier

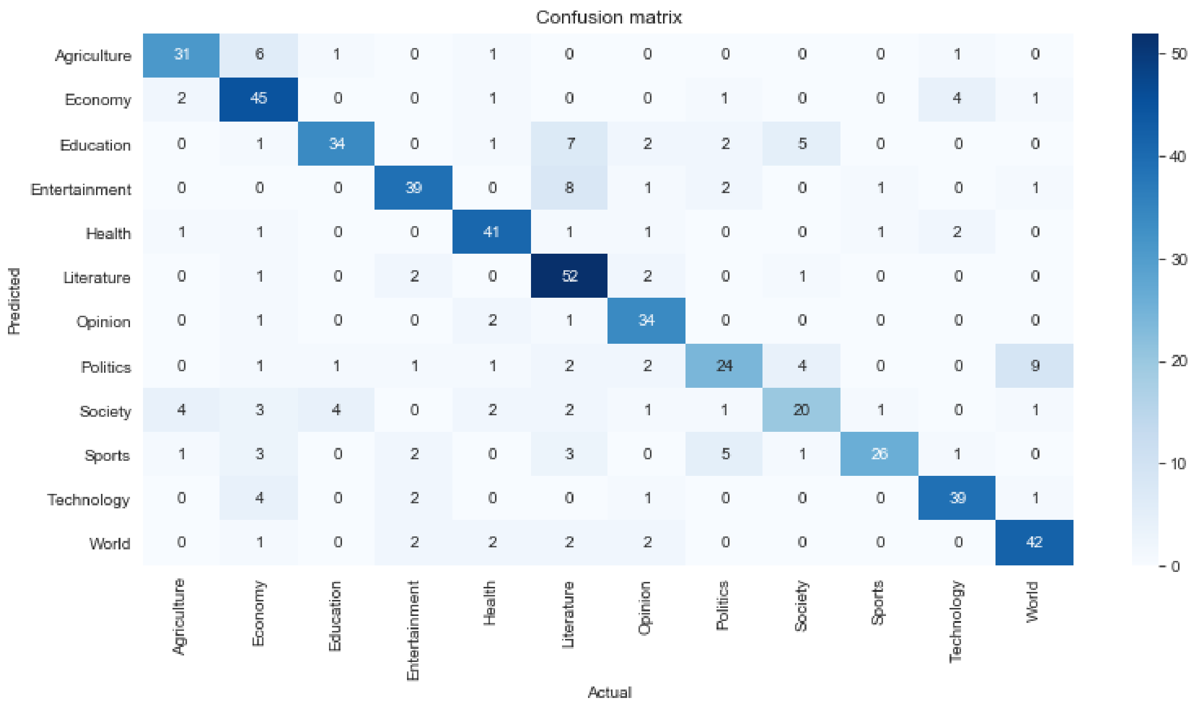
Categories	Test Samples	Predicted Correctly	Predicted Incorrectly
Agriculture	40	34	6
Education	51	38	13
Economy	55	43	12
Entertainment	43	40	3
Technology	43	38	5
World	54	46	8
Opinion	50	37	13
Politics	45	30	15
Society	31	21	10
Sports	41	31	10
Literature	67	53	14
Health	46	38	8

Table 4: Test Result for Logistic Regression Classifier

Therefore, Accuracy= $\frac{CorrectlyPredictedSamples}{TotalTestSamples}$

=449/566=0.793%

Naïve Bayes Classifier



	Agriculture	Economy	Education	Entertainment	Health	Literature	Opinion	Politics	Society	Sports	Technology	World
Agriculture	31	6	1	0	1	0	0	0	0	0	1	0
Economy	2	45	0	0	1	0	0	1	0	0	4	1
Education	0	1	34	0	1	7	2	2	5	0	0	0
Entertainment	0	0	0	39	0	8	1	2	0	1	0	1
Health	1	1	0	0	41	1	1	0	0	1	2	0
Literature	0	1	0	2	0	52	2	0	1	0	0	0
Opinion	0	1	0	0	2	1	34	0	0	0	0	0
Politics	0	1	1	1	1	2	2	24	4	0	0	9
Society	4	3	4	0	2	2	1	1	20	1	0	1
Sports	1	3	0	2	0	3	0	5	1	26	1	0
Technology	0	4	0	2	0	0	1	0	0	0	39	1
World	0	1	0	2	2	2	2	0	0	0	0	42

Figure 16: Confusion Matrix for Naïve Bayes Classifier

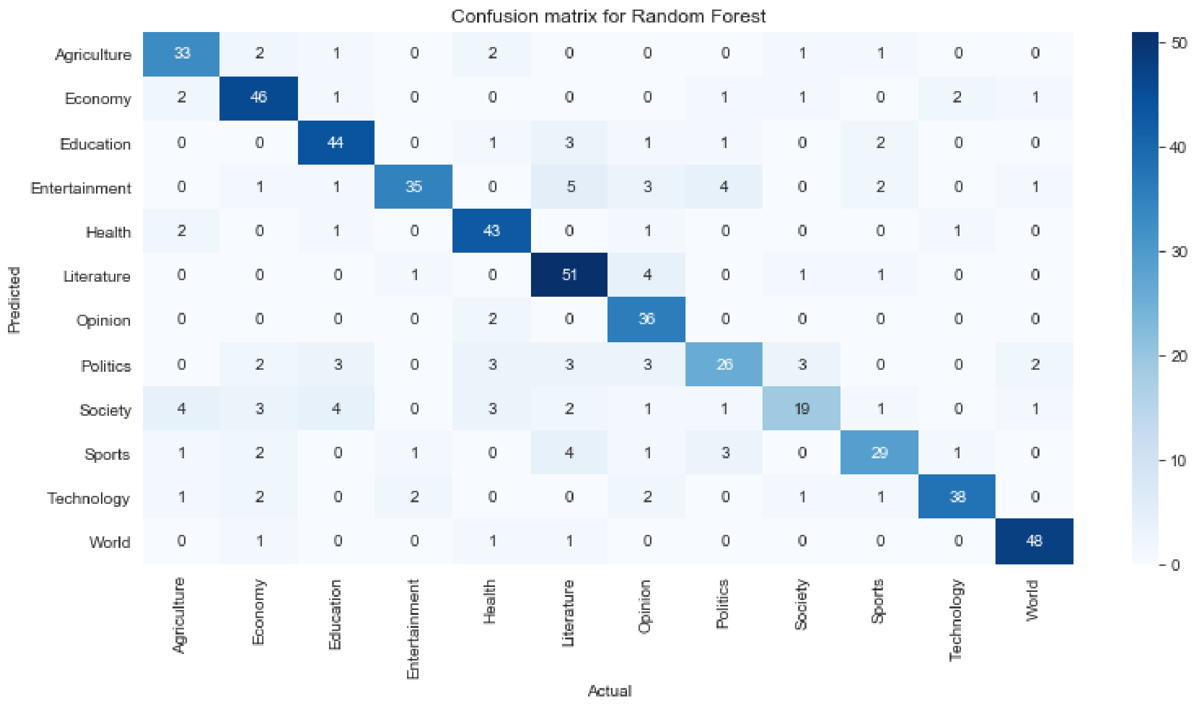
Categories	Test Samples	Predicted Correctly	Predicted Incorrectly
Agriculture	39	31	8
Education	40	34	6
Economy	67	45	22
Entertainment	48	39	9
Technology	47	39	8
World	55	42	13
Opinion	46	34	12
Politics	35	24	11
Society	31	20	11
Sports	29	26	3
Literature	78	52	26
Health	51	41	10

Table 5: Test Result for Naïve Bayes Classifier

Therefore, Accuracy= $\frac{CorrectlyPredictedSamples}{TotalTestSamples}$

=427/566=0.754%

Random Forest Classifier



	Agriculture	Economy	Education	Entertainment	Health	Literature	Opinion	Politics	Society	Sports	Technology	World
Agriculture	33	2	1	0	2	0	0	0	1	1	0	0
Economy	2	46	1	0	0	0	0	1	1	0	2	1
Education	0	0	44	0	1	3	1	1	0	2	0	0
Entertainment	0	1	1	35	0	5	3	4	0	2	0	1
Health	2	0	1	0	43	0	1	0	0	0	1	0
Literature	0	0	0	1	0	51	4	0	1	1	0	0
Opinion	0	0	0	0	2	0	36	0	0	0	0	0
Politics	0	2	3	0	3	3	3	26	3	0	0	2
Society	4	3	4	0	3	2	1	1	19	1	0	1
Sports	1	2	0	1	0	4	1	3	0	29	1	0
Technology	1	2	0	2	0	0	2	0	1	1	38	0
World	0	1	0	0	1	1	0	0	0	0	0	48

Figure 17: Confusion Matrix for Random Forest Classifier

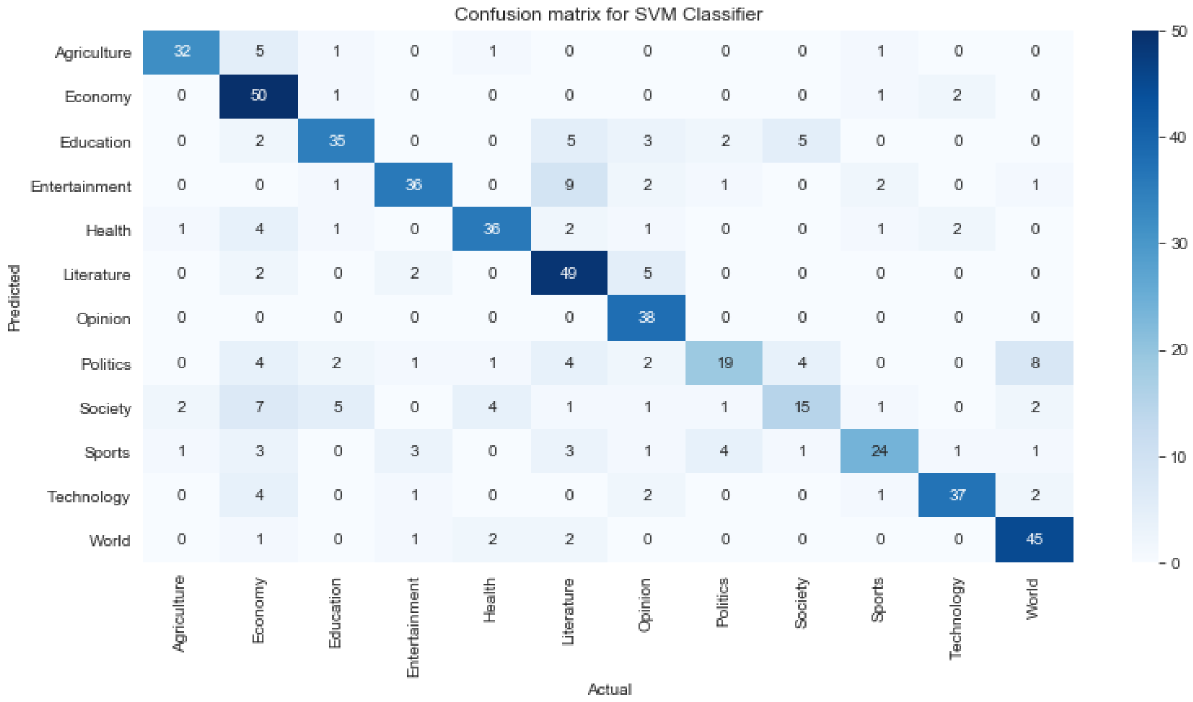
Categories	Test Samples	Predicted Correctly	Predicted Incorrectly
Agriculture	43	33	10
Education	55	44	11
Economy	59	46	13
Entertainment	39	35	4
Technology	42	38	4
World	53	48	5
Opinion	52	36	16
Politics	36	26	10
Society	26	19	7
Sports	37	29	8
Literature	69	51	18
Health	55	43	12

Table 6: Test Result for Random Forest Classifier

Therefore, Accuracy= $\frac{CorrectlyPredictedSamples}{TotalTestSamples}$

=448/566=0.791%

SVM Classifier



Predicted \ Actual	Agriculture	Economy	Education	Entertainment	Health	Literature	Opinion	Politics	Society	Sports	Technology	World
Agriculture	32	5	1	0	1	0	0	0	0	1	0	0
Economy	0	50	1	0	0	0	0	0	0	1	2	0
Education	0	2	35	0	0	5	3	2	5	0	0	0
Entertainment	0	0	1	36	0	9	2	1	0	2	0	1
Health	1	4	1	0	36	2	1	0	0	1	2	0
Literature	0	2	0	2	0	49	5	0	0	0	0	0
Opinion	0	0	0	0	0	0	38	0	0	0	0	0
Politics	0	4	2	1	1	4	2	19	4	0	0	8
Society	2	7	5	0	4	1	1	1	15	1	0	2
Sports	1	3	0	3	0	3	1	4	1	24	1	1
Technology	0	4	0	1	0	0	2	0	0	1	37	2
World	0	1	0	1	2	2	0	0	0	0	0	45

Figure 18: Confusion Matrix for SVM Classifier

Categories	Test Samples	Predicted Correctly	Predicted Incorrectly
Agriculture	36	32	4
Education	46	35	11
Economy	76	50	26
Entertainment	44	36	8
Technology	42	37	5
World	58	45	13
Opinion	55	38	17
Politics	27	19	8
Society	25	15	10
Sports	31	24	7
Literature	75	49	26
Health	44	36	8

Table 7: Test Result for SVM Classifier

Therefore, Accuracy= $\frac{CorrectlyPredictedSamples}{TotalTestSamples}$

=416/566=0.748%

Training and Test Set Accuracy Comparison

S.N.	Models	Training Set Accuracy	Test Set Accuracy
0	Gradient Boosting	0.99	0.775
1	KNN	0.99	0.701
2	Logistic Regression	0.87	0.793
3	Naïve Bayes	0.82	0.754
4	Random Forest	0.99	0.791
5	SVM	0.79	0.748

Table 8: Training and Test Set Accuracy

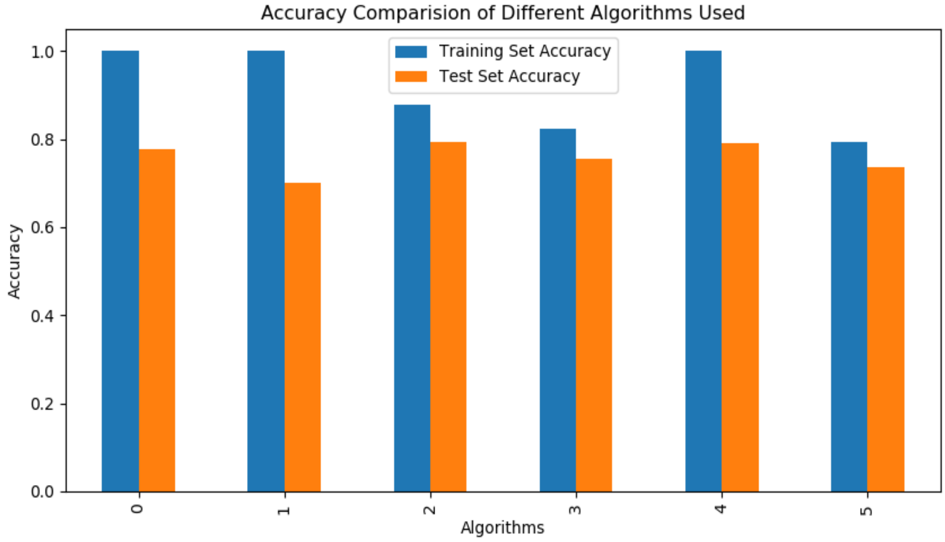


Figure 19: Training and Test Data Accuracy Comparison

The training dataset was splitted into train data and test data, keeping 20% (around 550 of 2800 articles) of total as test data. This was done so, to observe that how the models perform for untrained data. The result obtained is illustrated in the table (2) and figure (14) above, which shows that though very high accuracy was found in train data, comparatively less accuracy was in case of test data, which is obvious. The extent of variation in the accuracies in train and test data is bias and we found higher bias in context of models like- Gradient Boosting, KNN and Random forest than others illustrating these models to have overfitted. So, taking into consideration of overfitted models, Logistic Regression and SVM are comparatively better than other models in this project.

5. CONCLUSION AND FUTURE WORKS

5.1. Conclusion

The project entitled “Nepali News Classifier” conducted as a final year project of Bachelor Degree in Computer Engineering was completed successfully. The emphasis of this project was to study, analyze, compare and implement different classification algorithms and to develop a practical understanding of the domain of Natural Language Processing and Machine Learning.

The supervised classification used for the project were SVM, Naïve Bayes, kNN, Random Forest, Gradient Boosting and Logistic Regression. Among various matrices such as Confusion Matrix, Precision, Recall and Accuracy, this project focused on Accuracy, as major performance factor. The accuracy in training data is highest for SVM model for less amount of data at the first iteration, than for other models. For increased data, the accuracy increased and logistic regression models has highest accuracy than others. The test set accuracy differs to high extent from train set accuracy in case of Gradient Boosting, Random Forest and k-NN, causing these models overfitted. So, it is concluded that Logistic Regression Algorithm is better than others for Nepali News Classification Problem.

5.1. Future Work

The future recommendations will help other researcher interested in the domain of Nepali NLP and specially Classification problems, to provide more better and optimized solution for such problem. Some major recommendations are enlisted below:

- This project trained, around 2900 articles only and increasing training data will give more accurate models for classification.
- This project only considers 12 categories of Nepali News and adding other possible subjective categories will help to enhance the application to be more efficient, accurate and realistic.
- This application for News Classification can be implemented using web as well as mobile platforms like Android.

6. PROJECT TASK AND TIME SCHEDULE

The project schedule has been designed and followed as per requirements and constraints involved. This project is completed in about 2 months. We emphasized on *Data Collecting, Appropriate Dataset Formation, model training* and *Testing.* While carrying out this project, we documented every other thing to make sure for. Documentation is evidence of a good project development and its success. Training Models and Refining it, then Testing and Debugging should prioritize along with the fine documentation.

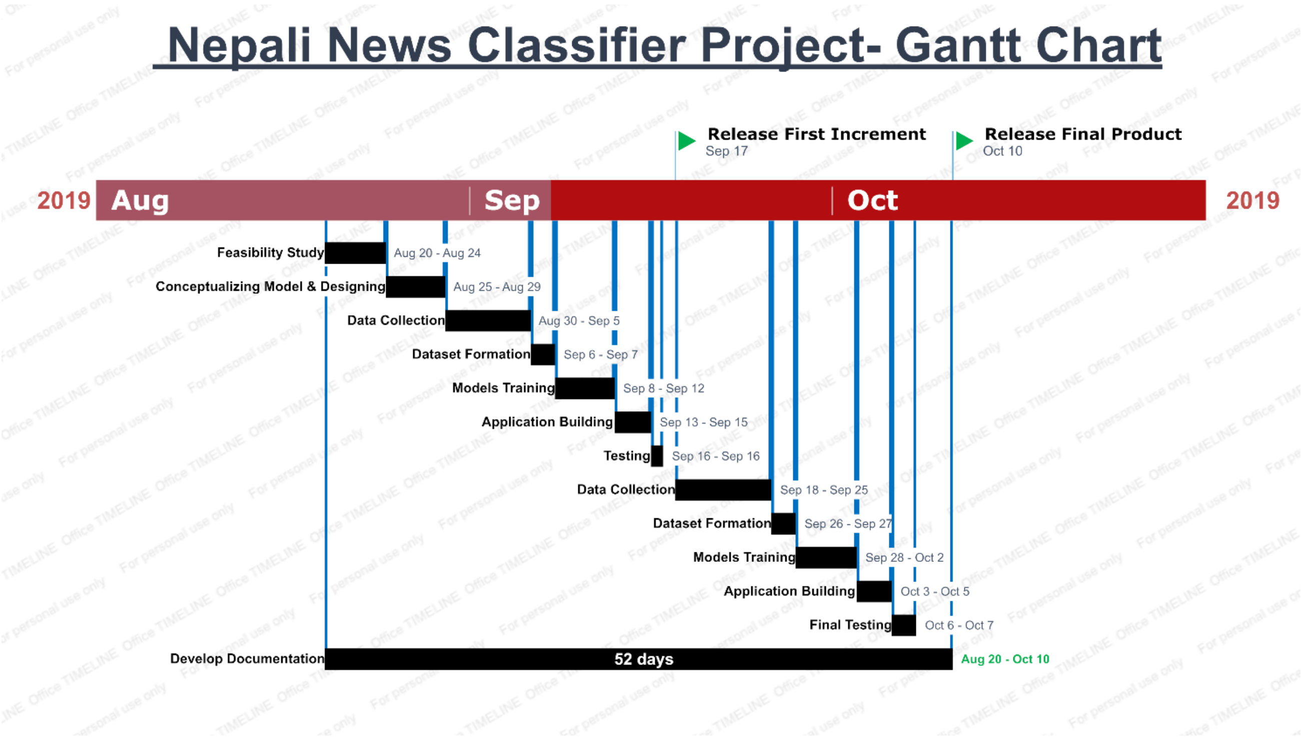


Figure 20: Gantt Chart

References

[1] Dr. Michael J. Garbade, A Simple Introduction to Natural Language Processing, April 2018, accessed on: 4 August, 2019

[2] Jean Mark Gawron, Combining Linguistic and Statistical Knowledge Sources in Natural Language Processing for ATIS, 12 March 1996, accessed on: 16 Sept, 2019

[3] Tej Bahadur Shahi; Ashok Kumar Pant, Nepali news classification using Naïve Bayes, Support Vector Machines and Neural Networks, IEEE *Xplore***:** 29 March 2018, accessed on:10 Aug 2019

[4] Supervised Learning for Text Classification, https://blog.thedigitalgroup.com/supervised learning-for-text-classification, , accessed on: 18 Sept 2019

[5] Nepali NLP Projects at Language Technology Kendra, http://ltk.org.np/projects.php, accessed on:15 Aug 2019

[6] Susan Li, Multi-Class Text Classification with Scikit-Learn, Feb 19, 2018, accessed on 26 Aug 2019

[7] Why the future of NLP is in the hands of tech giants ? , https://analyticsindiamag.com/why-the-future-of-nlp-is-in-the-hands-of-tech-giants-like-google-microsoft-amazon/, referenced on 18 Aug, 2019

[8] Y. H. Li and A. K. Jain, "Classification of Text Documents," [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.100.7400&rep=rep1&type=pdf , accessed on 16th Nov 2019

[9] Support Vector Machines for Text Categorization, A. Basu, C. Watters, and M. Shepherd, 36th Hawaii International Conference on System Sciences,2002, IEEE, accessed on 20th Nov 2019

[10] Thorsten Joachims, paper on "Text Categorization with Support Vector Machines: Learning along with many other language relevant features", [Online] Available: http://www.cs.cornell.edu/people/tj/publications/joachims98a.pdf, accessed on 18 Nov, 2019

[11] Nepali news classification using Naïve Bayes and Support Vector Machines
Conference: 2018 International Conference on Communication information and Computing Technology (ICCICT), February 2018, accessed on 20th Nov, 2019

[12] Tej Bahadur Shahi and Abhimanu Yadav. Mobile SMS spam filtering for Nepali text using naive Bayesian and support vector machine. International Journal of Intelligence Science, 4(01):24, 2013, accessed on 12th Dec, 2019

[13] Dinesh Dangol and Arun K. Timalsina. Effect of Nepali language features on Nepali news classification using vector space model. 2013, accessed on 8th Nov,2019

[14] Surya Bahadur Bam and Tej Bahadur Shahi. Named entity recognition for Nepali text using support vector machines. Intelligent Information Management, 2014,2014., accessed on 11th Nov, 2019

[15] Tkinter GUI tutorial video, https://www.youtube.com/watch?v=KKpf0EcgkUU, referenced on: 22 Aug 2019

[16] Support Vector Machine, https://en.wikipedia.org/wiki/Support-vector_machine, accessed on 5th Oct,2019

[17] Scikit Learn Package Documentation, https://scikit-learn.org/stable/ referenced on: 24 July 2019 and later many times.

[18] Web Scrapping References, https://towardsdatascience.com/web-scraping-news-articles-in-python-9dd605799558, referenced on: 16 Aug 2019

[19] NCIT Project Guidelines and Sample Reports, Previous Projects Documents https://drive.google.com/drive/folders/0By46xMRiORSmS05EZEdsaTZGNzQ, referenced on: 20 July 2019 and several times.