TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

THAPATHALI CAMPUS

A Minor Project Proposal

On

Tender Notice extraction from E-papers using Neural Network

Submitted By:

Anuj Sedhai	(THA075BCT010)
Ashmin Bhattarai	(THA075BCT012)
Devraj Neupane	(THA075BCT019)
Manish Khadka	(THA075BCT025)

Submitted To:

Department of Electronics and Computer Engineering

Thapathali Campus

Kathmandu, Nepal

December, 2021

ABSTRACT

Tender notices are very much important to financial institutions like construction companies, contractors as well as government institutions. These notices consist of all the required information about the tender which are often published in daily as well as weekly newspapers. However, it is very much tedious for these companies to manually search tender notices in every newspaper and figure out which bid is best suited for them. This project is built with the purpose of solving this tedious task of manually extracting the tender notices. The e-papers will be scanned and tender notices are automatically extracted using neural network. For extraction we will train SVM and different architectures of CNN and model with highest accuracy will be implemented. These extracted notices will be then published in website. This project might be helpful for construction companies as well as contractors assuring quality and efficiency. This project has great application in the field of competitive bidding as well as managing them in a systematic manner.

Keywords: E-papers, Neural Network, Tender Notices, CNN, Website

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
CNN	Convolution Neural Network
CUDA	Compute Unified Device Architecture
et al	And Others
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
I/O	Input Output
ML	Machine Learning
NN	Neural Network
ResNet	Residual Network
RFT	Request For Tender
ROI	Region of Interest
SVM	Support Vector Machine
UI	User Interface

1. INTRODUCTION

1.1. Background

A tender is an invitation to bid for a project or accept a formal offer such as a takeover bid. Tendering usually refers to the process whereby governments and financial institutions invite bids for large projects that must be submitted within a finite deadline.

For projects or procurements, most institutions have a well-defined tender process, as well as processes to govern the opening, evaluation, and final selection of the vendors. This ensures that the selection process is fair and transparent. When it comes to tender offers for takeover attempts, the conditions of the offer are clearly listed and include the purchase price, the number of shares asked, and a deadline for a response.

A request for tender (RFT) is a formal and structured invitation to suppliers to submit competitive bids to supply raw materials, products, or services. Because this is a public and open process, laws were created to govern the process to ensure fair competition among bidders.

In the context of Nepal, tender is generally published in the newspaper. But as the new generation is becoming more connected to the digital world, the tender notice published in the Newspaper becomes less effective.

AI, the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. The term is often applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from experience. Since the development of the digital computer in the 1940s, it has been proved that computers can be programmed to carry out overly complex tasks. Still, despite continuing advances in computer processing speed and memory ability, there are yet no programs that can match human flexibility over wider domains or in tasks requiring much everyday knowledge. On the other hand, some programs have reached

the performance levels of human experts and professionals in performing certain specific tasks. Artificial intelligence can be found being widely used in the field of medicine, computer search engines, transportation and so on.

1.2. Motivation

There are several tender portals in Nepal which provide tender notices published in newspaper. It is done by clipping the notice from e-papers of respective newspaper one by one and publishing them in the website. For these purposes, these companies usually hire a team of 2-3 people to identify, crop and publish that tender notice to their respective tender portal. However, employing people can be costly as well as there might be chances of missing the notices because human beings are prone to make errors. Considering these factors, we came up with the idea of automating the entire process using artificial intelligence. This can be helpful in achieving efficiency as well as reducing the expenses.

1.3. Problem Definition

The problem in the project is to scan the e-papers of national as well as local newspapers and find the tender notices using artificial intelligence. The program will be able to distinguish between notices and news articles by identifying some unique attributes of tender notices.

1.4. Project Objectives

The main objectives of our project are listed below**:**

- To extract tender notices from newspapers using neural network
- To publish extracted notices in a website

1.5. Project Scope and Application

The application of our project is to gather Tender notice from different e-papers and publish them in a website, so that the customer doesn't need to read individual newspaper to find all the tender notices published on that day. This project can be very

much useful for big contractors and construction company as it can save their time, efforts and cost as well. Using this system, it can be very much easier for them to bid in competitive biddings, gather all the information about the tender, all in a single platform.

2. LITERATURE REVIEW

This project uses a machine learning algorithm. The big data and machine learning technologies can be used for econometrics enterprises, tender evaluation, or analysis of public procurement notices.

There is extensive literature about tender evaluation (also called bidding selection methods) for the selection of the best supplier in public procurement with different techniques such as the economic scoring formulas, data envelopment analysis or multicriteria decision making, and where multiple bidders are evaluated based on price and quality. In particular, the most studied public procurement auctions are related to construction, i.e., distribution of bids, bidding competitiveness and position performance, strategic bidding, tender evaluation and contractor choice.

Another relevant subject in the public procurement literature is the detection of collusive tendering or bid rigging with case studies in Spain, India, and Hungary. This occurs when businesses that would otherwise be expected to compete secretly conspire to raise prices or lower the quality of goods or services for purchasers in a public procurement auction (this is called a cartel). In addition, public procurement contracts have other issues such as best quality, too many regulations, systemic risk, or corruption. Corruption is a form of dishonesty undertaken by a person or organization with the authority to get illicit benefit. There are empirical studies to detect corruption by analyzing public tenders in many countries, for example, in China, Russia, the Czech Republic, and Hungary. The application of algorithms by governments or enterprises to detect collusion or corruption, especially using machine learning methods, has become an almost inevitable topic and the subject of many studies.

2.1. Related works:

Manuel J. García Rodríguez, et. al [1] developed bidders recommender for public procurement auctions which recommend potential bidders using machine learning particularly random forest classifier. In this project, the winning bidding company was within the recommended companies group, from 24% to 38% of the tenders, according to different test conditions and scenarios. However, it is described theoretically and

validated experimentally using a case study from Spain and can be implemented or adapted at any particular situation

Wei-Ta Chu, et. al [2] proposed advertisement detection and segmentation where they detected advertisement candidates based on a connected components method and use CNN with SVM to classify advertisement from rest of the contents. However, due to insufficient dataset, they achieved average 85.54% accuracy using only CNN, 84.65% using only semantics and 94.04% using both.

Pooja Jain, et. al [3] from Punjab university, Department of Science and application developed a convolutional neural network-based advertisement classification models for online newspapers that can classify advertisement images from English newspapers into four pre-defined categories. Initially, they used simple CNN-model for advertisement classification purpose which in turn gave an accuracy of 65%. Similarly, they used ResNet50 architecture along with default hyperparameters and achieved accuracy of 68%. Finally, implementation of ResNet50 architecture with Finetuned hyperparameters yield an accuracy of 74%.

Kim Sungyoung, et. al.[4] proposed a method for extracting interesting object from complex background . In this method, a core object region was selected as a region a lot of pixels of which had the significant color, and then it was grown by iteratively merging its neighbor regions and ignoring background regions. But in this project some inaccuracies occurred because of the under-extraction or over-extension of wrong significant regions

An Tien Vo, et. al. [5] proposed image classification model that can be applied for identifying the display of online advertisement. They used convolutional neural network with two parameters (n, m) where n is a number of layers and m is number of filters in Conv layer. However, using Convolutional Neural Network only extracts image advertisement only but advertisements might be published in textual form as well. In such scenarios, this model might be inefficient.

Generally tender notices in E-Papers are published in textual form i.e. newspaper themselves publish the notice on the behalf of contractors. But in some newspapers

contractors themselves provide notice in image form to the newspapers. Hence in this project we have considered image processing as well as OCR.

3. METHODOLOGY

This entire project is composed of two phases: training phase and implementation phase. First, we collect data needed for this project. Then we train suitable Neural Network model. After training a model it is used in implementation phase where we separate out tender advertisement from reset of news articles.

3.1. Elaboration of Working Principal

3.1.1. Development of Training Models

The datasets for training model are collected from employees of those companies who are doing this task manually. Furthermore e-newspapers are collected form their respective website. To collect huge amount of news we will use help of web scrapping library: Beautifulsoup available in python. Once we collect all required data then another step is to filter data i.e., separation of news/advertisements from rest of the news articles. Since it is always the case that ads on newspapers are in proper format and contained inside a rectangular box. Hence with the use of OpenCV library we will detect contours in a page of newspaper. By setting certain threshold value we can avoided some small noises while detecting contours. After getting ROI we stack all of it in a panda's data Frame [6]. Similarly, all data collected previously from employees are also stacked on data frame, which will be positive image for our training datasets. For negatives datasets images are collected from google image.

Now we will enter into training phase. Here will make several models with different hyperparameters and will train all models. The model which gives most accuracy in validation phase, that will be our final model. For training we will use Keras API [7] inside TensorFlow library of python. Models and architecture used are describe below.

3.1.1.1. Support Vector Machine

A Support Vector Machine (SVM) is a powerful and versatile Machine Learning model, capable of performing linear or nonlinear classification, regression and outlier detection. In case of two-dimensional dataset SMV will find best line to separate out two or more classes. Similarly, for multi-dimensional dataset it will find best fitting

plane. SVM will give best result if the dataset is very small. But if there are large number of features to be taken into account the SVM will comparatively give lower accuracy.

3.1.1.2. Convolution Neural Network

Convolutional neural networks (CNNs) emerged from the study of the brain's visual cortex, and they have been used in image recognition since the 1980s. In the last few years, thanks to the increase in computational power, the amount of available training data, CNNs have managed to achieve superhuman performance on some complex visual tasks. They power image search services, self-driving cars, automatic video classification systems and many more. CNN is composed of convolution layer pooling layer and dense layer. First image is fed into convolution layer of certain filter size, which apply filters and extract certain features from images. Then it is sent to pooling layer where spatial size of images is reduces without losing features inside image. Different approach has taken in different architecture for stacking convolution and pooling layer. Moreover, there is a layer called activation function in between convolution and pooling layer. At last, there are some layers of dense network. The numbers of neurons in output layer are defined by type of classification whether it is binary or multi-class classification. Some common architecture used are mention below.

3.1.1.2.1. GoogLeNet

The GoogLeNet architecture was developed by Christian Szeged et al. from Google Research,13 and it won the ILSVRC 2014 challenge by pushing the top-five error rate below 7%. This great performance came in large part from the fact that the network was much deeper than previous CNNs. This was made possible by subnetworks called inception modules which allows GoogLeNet to use parameters much more efficiently.

3.1.1.2.2. ResNet

Kaiming He et al. won the ILSVRC 2015 challenge using a Residual Network (or ResNet), that delivered an astounding top-five error rate under 3.6%. The winning variant used an extremely deep CNN composed of 152 layers (other variants had 34, 50, and 101 layers). The key to being able to train such a deep network is to use skip

connections (also called shortcut connections): the signal feeding into a layer is also added to the output of a layer located a bit higher up the stack. When training a neural network, the goal is to make it model a target function h(x). If we add the input x to the output of the network (i.e., we add a skip connection), then the network will be forced to model f(x) = h(x) – x rather than h(x). This is called residual learning.

3.1.1.2.3. Xception

Xception is a variant of GoogLeNet architecture which stands for Extreme Inception. It was proposed in 2016 by Francois Chollet. It merges the idea of GoogLeNet and ResNet, but replaces the inception modules with a special type of layer called a depthwise separable convolution layer. While a regular convolutional layer uses filters that try to simultaneously capture spatial patterns (e.g., an oval) and cross-channel patterns (e.g., mouth + nose + eyes = face), a separable convolutional layer makes the strong assumption that spatial patterns and cross-channel patterns can be modeled separately.

3.1.2. Implementing Trained Model

In this phase we will use two models one for separating advertisement and another for recognizing tender notice from rest of notices. For separating we will use model we just trained on training phase. For extraction of tender notice, due to time limitation we will use Tesseract model which recognizes text in images. Hence finding keywords related to tender like Invitation for Bids, Request for Proposal, Expression of Interest, Environmental Impact Assessment (EIA), Letter of Intent, Procurement Notice, Auction, etc. After separating tender notice, it will be published in a very simple website.

3.1.3. Blocked Diagram

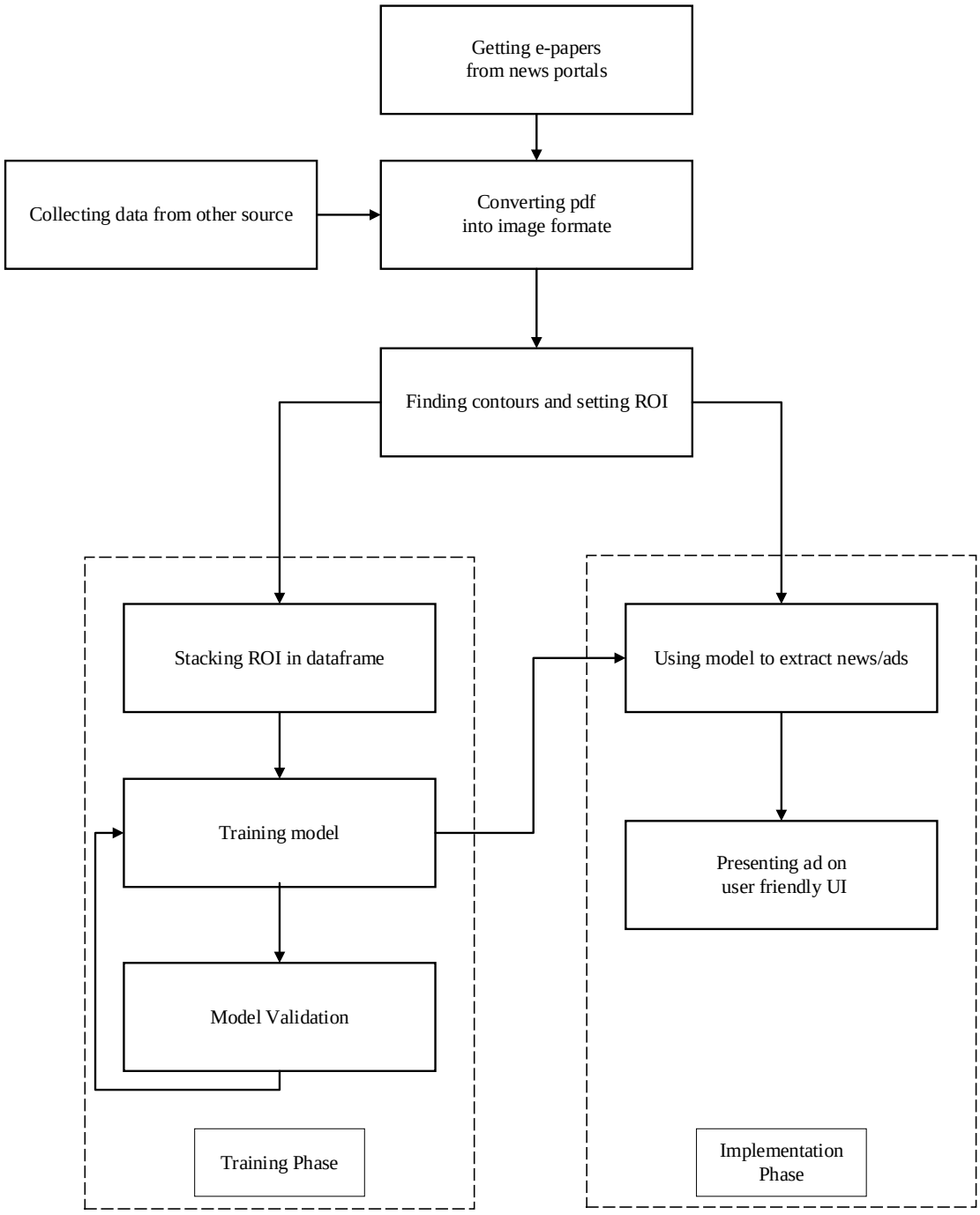


Fig 3.1. Proposed System Block Diagram

3.2. Dataset

There is lack of standard datasets of tender notices from E-Papers and therefore we created our own dataset from employees working in the tender portal company named Bolpatra Nepal. A balanced dataset of approx. 15000 positive images and 15000 negative images are collected and approximately 5000 images are created using images augmentation.

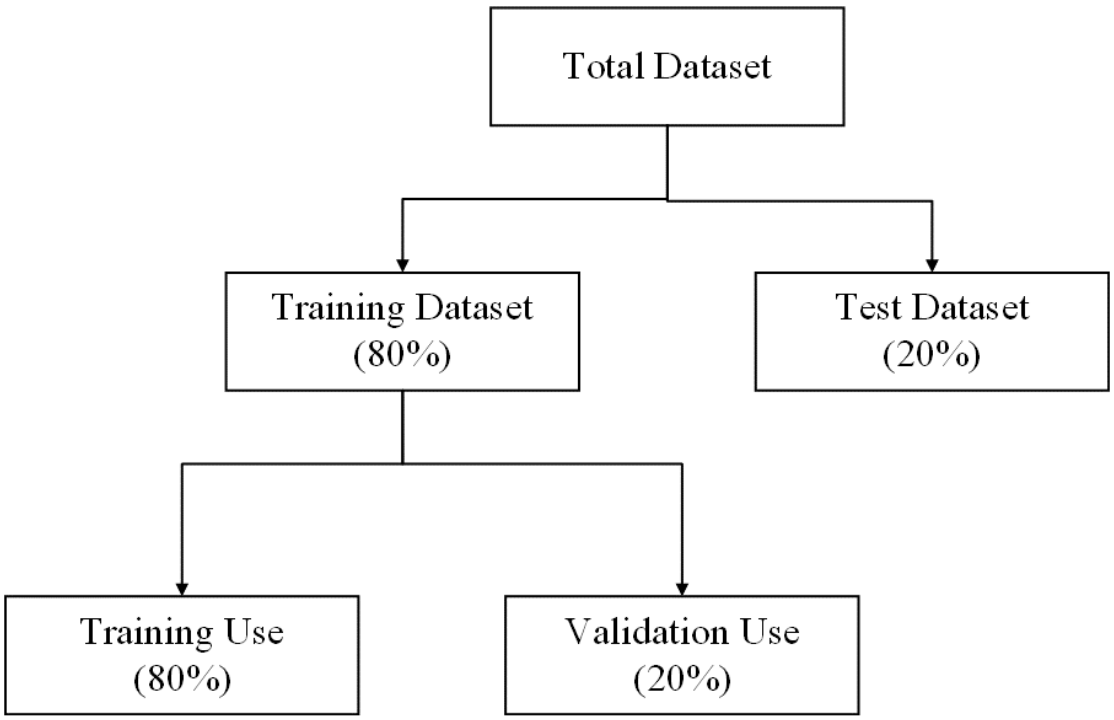


Fig 3.2. Division of Dataset

3.3. Hardware Requirement

3.3.1. CPU

Since this project involves lots of computation, CPU must be atleast intel i5 with 4 cores or equivalent.

3.3.2. RAM

As AI is major aspect of this project and to make a good AI model, we need significant amount of data. But is not the case that always data is in required format. Hence, we need to do some preprocessing for data and we need good space in RAM. Thus, RAM with 16GB space may works better.

3.3.3. GPU

As training a model is repetitive task and lots of core is needed for parallel computing. A single CPU can't provide huge amount of core. Thus, GPU having more than 1500 CUDA cores and atleast 6 GB of VRAM is mandatory.

3.4. Software Requirement

3.4.1. IDE and Text Editor

3.4.1.1. VSCode

Visual Studio Code is a source-code editor made by Microsoft for Windows, Linux and macOS. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git.

3.4.1.2. Jupyter Notebook

Jupyter Notebook is a web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning.

3.4.2. Language

3.4.2.1. Python

Python is a high-level scripting language which can be used for a wide variety of text processing, system administration and internet-related tasks. Unlike many similar languages, its core language is very small and easy to master, while allowing the addition of modules to perform a virtually limitless variety of tasks. Python is a true object-oriented language, and is available on a wide variety of platforms. There's even a python interpreter written entirely in Java, further enhancing python's position as an

excellent solution for internet-based problems. Python version 3.9 is mandatory for this project.

3.4.3. Libraries

3.4.3.1. NumPy

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

3.4.3.2. Pandas

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. The name Pandas is derived from the word Panel Data – an Econometrics from Multidimensional data. In 2008, developer Wes McKinney started developing pandas when in need of high performance, flexible tool for analysis of data.

3.4.3.3. Matplotlib

Matplotlib is a multiplatform data visualization library built on NumPy arrays, and designed to work with the broader SciPy stack. It was conceived by John Hunter in 2002, originally as a patch to IPython for enabling interactive MATLAB-style plotting via gnuplot from the IPython command line.

3.4.3.4. BeautifulSoup

Beautiful Soup is a library that makes it easy to scrape information from web pages. It sits atop an HTML or XML parser, providing Pythonic idioms for iterating, searching, and modifying the parse tree.

3.4.3.5. OpenCV

OpenCV (Open-Source Computer Vision Library) is a library of programming functions mainly aimed at real-time computer vision. Originally developed by Intel, it was later supported by Willow Garage then Iteses (which was later acquired by Intel). The library is cross-platform and free for use under the open-source Apache 2 License. Starting with 2011, OpenCV features GPU acceleration for real-time operations.

3.4.3.6. TensorFlow

TensorFlow is a Python library for fast numerical computing created and released by Google. It is a foundation library that can be used to create Deep Learning models directly or by using wrapper libraries that simplify the process built on top of TensorFlow. TensorFlow version 2.11 is compatible with python version 3.9.

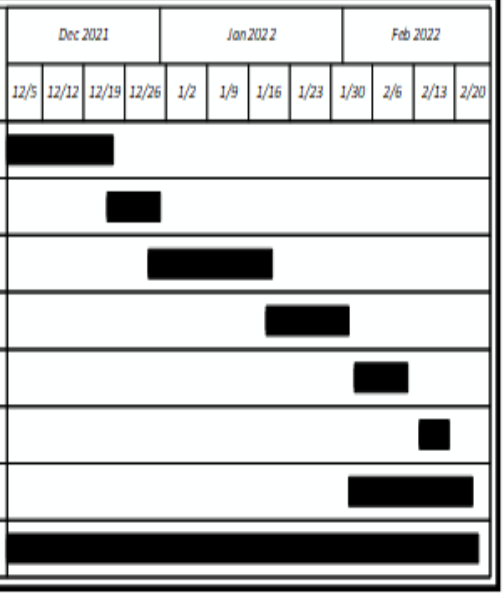
3.4.3.6. Keras

Keras is a high-level Deep Learning API that allows to easily build, train, evaluate and execute all sorts of neural networks. It was developed by Francois Chollet as part of a research project and was released as an open-source project in March 2015.

4. ESTIMATE PROJECT SCHEDULE

Table 4. 1. Gantt Chart

ID	Task Name	Start	Finish	Duration
1	Research	12/6/2021	12/23/2021	14d
2	Data Collection	12/23/2021	12/31/2021	7d
3	Data Filtering	12/30/2021	1/19/2022	15d
4	Training	1/19/2022	2/1/2022	10d
5	Designing UI	2/3/2022	2/11/2022	7d
6	Integrating Model with UI	2/14/2022	2/18/2022	5d
7	Testing and Dubugging	2/2/2022	2/22/2022	15d
8	Documentation	12/6/2021	2/23/2022	58d



5. EXPECTED OUTPUT

After the completion of this project, we expect our program to scan all the pages of E-papers of respective newspapers and identify tender notices. Later on, these extracted notices can be published in a user-friendly website. In the website, the user can easily access all the related and important information regarding the procurement notice in which he/she wishes to bid.

6 FEASIBILITY ANALYSIS

6.1 Economic Feasibility

The important economic factor for this project is dataset and we created a manual dataset using images collected from employees of tender portal company. Hence, this project is economically feasible.

6.2 Technical Feasibility

In this project we have used convolutional neural network as well as support vector machine which require computer system with high computational power. Therefore, this project isn't suitable for developers having lower end computer setup.

6.3 Operational Feasibility

The extracted notices are collected and published in a simple and user-friendly website for which user doesn't require any programming background. The user can simply visit the site and explore the desired notices. Hence, it is operationally feasible.

6.4 Legal Feasibility

For this project we have been granted approval from Bolpatra Nepal in order to use datasets collected from employee working in that company. Hence, the project is legally feasible.

REFERENCES

[1] Manuel J. García Rodríguez, Vicente Rodríguez Montequín, Francisco Ortega Fernández, Joaquín M. Villanueva Balsera, "Bidders Recommender for Public Procurement Auctions Using Machine Learning: Data Analysis, Algorithm, and Case Study with Tenders from Spain", *Complexity*, vol. 2020, Article ID 8858258, 20 pages, 2020.

[2] Wei-Ta Chu and Han-Yuan Chang, "Advertisement Detection, Segmentation, and Classification for Newspaper Images and Website Snapshots," Proceedings of International Computer Symposium, 2016

[3] Pooja Jain, et. al. "Convolutional Neural Network Based Advertisement Classification Models for Online English Newspapers." (2021).

[4] Kim, Sungyoung & Park, Soyoun & Kim, Minhwan. (2003). Central Object Extraction for Object-Based Image Retrieval. Image and Video Retrieval. 50. 523-528. 10.1007/3-540-45113-7_5.

[5] A.T. Vo, H.S. Tran and T.H. Le, "Advertisement image classification using convolutional neural network," 2017 9th International Conference on Knowledge and Systems Engineering (KSE), 2017, pp. 197-202, doi: 10.1109/KSE. 2017.8119458.

[6] Vanderplas, J., 2016. *Python Data Science Handbook*. 1st ed. Sebastopol, CA: O'Reilly Media.

[7] Géron, A., 2019. *Hands-on machine learning with Scikit-Learn and TensorFlow*. 2nd ed. Sebastopol, CA: O'Reilly