PROJECT REPORT

CHARITY RECOMMENDATION ENGINE FOR WEB ARTICLES

delivered by

Kanchan Bhale (C22018111145)
Esha Sali (C22018111085)
Ritika Rana (C22018111080)

in partial fulfilment for the award of the degree of

Bachelor of Technology in

ELECTRONICS AND TELECOMMUNICATION of
SAVITRIBAI PHULE PUNEUNIVERSITY,

under the guidance of
Dr. Mrudul Dixit

Sponsored by : - Givetastic, Germany

in the Department of Electronics and Telecommunication of CUMMINS COLLEGE OF ENGINEERING FOR WOMEN, KARVENAGAR, PUNE -411052 (An Autonomous Institute affiliated to SAVITRIBAI PHULE PUNE UNIVERSITY)

Academic year
2021 - 22

a) **Project title: -** Charity Recommendation Engine For Web Articles

b) **Subject area: -** AI / ML

c) **Nature of the Project: -** Software

CERTIFICATE

This is to certify that

Kanchan Bhale

Esha Sali

Ritika Rana

have completed their Project on the Topic

Charity Recommendation Engine for Web Articles

in partial fulfilment for the award of the degree of

Bachelor of Technology in ELECTRONICS AND TELECOMMUNICATION

of SAVITRIBAI PHULE PUNE UNIVERSITY,

in

CUMMINS COLLEGE OF ENGINEERING FOR WOMEN, KARVENAGAR, PUNE-52

(An Autonomous Institute affiliated to SAVITRIBAI PHULE PUNE UNIVERSITY)

Internal Guide	Head of the Department	Principal
[Dr. Mrudul Dixit]	[Dr.Prachi Mukherji]	[Dr.M.B.Khambete]

Sponsorship Letter

Givetastic
Giving. Made Fantastic!

To,
Dr. Prachi Mukherjee
HOD, Electronics and Telecommunications Dept
Cummins College of Engineering for Women
Karvenagar, Pune - 411052

Vidya Munde-Müller
Givetastic Technologies UG
(haftungsbeschränkt)

Walderseestraße 7
30163 Hannover

09.11.2021

Subject: Sponsorship letter for B.Tech Project.

Dear Madam,

This letter is to confirm that our company, Givetastic Technologies, Germany will be sponsoring technical guidance on a project that three students namely:

1. Kanchan Bhale (C22018111145) - 4203
2. Esha Sali (C22018111085) - 4250
3. Ritika Rana (C22018111080) - 4242

from Final Year, B.Tech, Cummins College of Engineering for Women, Pune will be pursuing as Project Interns with Givetastic under the guidance of Mrs. Vidya Munde-Müller, CEO Givetastic.

Title of the project: Charity Recommendation Engine Based on Web Articles Read by User

Subject area of the project: AI / ML / Analytics

Nature of the project: Software Dev.

Vidya Munde-Müller

CEO & Founder

Givetastic Technologies UG

GIVETASTIC TECHNOLOGIES UG (HAFTUNGSBESCHRÄNKT)
VIDYA MUNDE-MÜLLER CEO
HOLVI BANK BIC: HOLVDEB1 DE25 1001 7997 0273 3804 72
Tel: +49 171 64 25 674 www.givetastic.org
AMTSGERICHT HANNOVER HRB 219237 USt-IdNr: DE327710533

Completion Certificate

Acknowledgement

First and foremost, we would like to express our gratitude towards our project guide Dr. Mrudul Dixit for her continuous support, patience and motivation throughout the completion of this project.

We would like to thank Givetastic, Germany for sponsoring our project and helping us in every possible way. Mrs. Vidya Munde-Müller (CEO) and Avaré Stewart (Technical Guide) have constantly helped us with several aspects of the project, be it managerial or technical. Our sincere thanks to them for their precious time.

We would also like to extend our sincere gratitude to Dr. Prachi Mukherjee (E&TC-HOD) for her valuable guidance, insightful feedback that pushed us to think out of the box.

Our thanks to the complete E&TC department for providing a platform for us to implement our project. We thank the teaching and non-teaching staff of Cummins College of Engineering for Women, Pune for their support throughout.

Kanchan Bhale (C22018111145)
Esha Sali (C22018111085)
Ritika Rana (C22018111080)

Abstract

Social Responsibility is a fundamental necessity in today’s world to fight social-evils. When it comes to Charity Donations, users face a lack of motivation due to the process being difficult and tedious. There is an unavailability of a one-stop platform where users can get recommendations of multiple trusted Charities based on articles read by them online. As for NGOs, there is a need of expanding their reach and influence.

The Recommendation Engine takes Web-Articles as input, pre-processes the raw text data using Natural Language Processing. For Classification, we have tested 5 different algorithms- Naïve Bayes, Logistic Regression, Support Vector Machines, Stochastic Gradient Descent and Random Forest algorithms. It then predicts the Social Cause highlighted in the article and recommends top 3 charities based on the predicted cause and various other metrics like user-ratings, financial-ratings, etc. This Recommendation Engine is integrated with a Web-Application built using Streamlit, an open-source Python framework. Streamlit builds tunnel websites that can be accessed as long as the Python Notebook runs online.

The scope of this project can be extended to provide User Customization based on location in order to assist the local NGOs with more exposure. Recommendations based on previous donations can also be provided. The developed Web-Application can be used to build a Web-Browser plugin, in the future.

Index

List of Figures

List of Tables

1. Introduction

We as humans have developed a sense of social responsibility, but In today's world Social Responsibility is not only a virtue but also a necessity. The main aim of this project is to empower individuals and NGOs to take immediate action on societal issues they care about and create a positive social impact. The goal of the AI model stated in this report is to encourage Social Responsibility through a Recommendation Engine that enables Users to choose the right charities for donation and help NGOs in their Reach and Engagement.

Individuals don’t know how easily they can contribute online to the causes they care about and Users face a lack of motivation to contribute for social causes and they often have to go through a tedious process before selecting an NGO of their choice. Also, NGOs need to reach and engage people online who are willing to support their cause and local NGOs require a platform through which they can expand their scope and influence. The AI model aims to bridge the gap between the user and the NGOs by developing a one stop platform that provides a list of reliable NGOs across the world and a non-biased and semantically accurate recommendation engine for end-users.

This model works as an end to end data science model that recommends charities based on the web articles read by the user on the internet. The design of this AI model will help classify the text, extract keywords and recognize different patterns present in the article read by the user. Developing this model will generate plug-ins that will analyze the text and help connect people to different charities for making donations based on the content they are reading on the internet.

1.1 Objective:

The overall objective behind the prototype development is to model a system that can be easily integrated on any news/literature or magazine article on certain web pages on the browser or with the help of mobile widgets, when the user scrolls through any sensitive/ charity/awareness or even a general article.

1.2 Goals:

Developing a traffic generation engine for non-profits and NGOs and building a non-biased and semantically accurate recommendation engine for end users. From user perspective, the model should also be able to provide customization based on location, preference, etc. while recommending charities. Along With this the main goal will be working on the integration of the plugin and AI based recommendation model.

1.3 Approach:

The model is trained on the Charity Navigator dataset which contains the category, cause, mission and tagline of various NGOs. While building, the model is trained and tested by using 5 different machine learning classifiers: Naive Bayes, Support Vector Machines, Logistic Regression, Random Forest and Stochastic Gradient Descent. Further the performance metrics of all these classifiers are compared using confusion matrix and classification report in order to select the best classifier. In the next step, classifier models are tested on random web articles (i.e. completely unknown data) as another performance metric. In order to improve the model efficiency, keywords and synonyms are extracted from the test document (random web article) and mapped these results with the categories(cause supported by the NGO) learned by the model from the Charity Navigator Dataset. With this approach, the model correctly predicts the cause supported by the web article which is read by the user on the internet and recommends NGOs according to the cause predicted as a result.

2. Literature Survey

Our literature survey is based on the pre-existing products and on various Journal and research papers that we read in order to implement our project.

A) Pre-existing Products

Table 2.1 Pre-existing Products

No.	Name of the existing product	Specifications and Features	In-use	Cost	Website
1.	WPForm	Drag & drop form builder for an existing website, customizable donation forms, one-time donations, survey forms, polls, volunteer signup forms etc. Easy to integrate with all major email marketing services	Used by over 4 million sites	Discounted license for non-profits at $99/year (75% off their regular price)	https://wpforms.com/
2.	GiveWP	One-time donations, google analytics integrations, fine-tune features like customizable donations tributes like “in honour of”, currency switcher options, incentives, etc		Basic paid plan starts at $240/year	https://givewp.com/ref/477/?campaign=wpbeginner

3.	Paypal	WordPress plugin to connect PayPal account with WordPress site, recurring monthly donation options.		One time donations - free, recurring donations : $59.95/year	https://wordpress.org/plugins/easy-paypal-donation/
4.	Woocommerce	Multiple levels of donations, set minimum and maximum levels for payment, specify a fixed amount for donation, or let the customers specify an amount of their liking, place the option on the product, cart or Checkout pages, or use a widget.		$99	https://woocom merce.com/pro ducts/donati on product-for-wo ocommerce/ ?af f=4166
5.	Tab for a cause	Every time you open a new browser tab, raise money for charity	3,000 users & don-ated $4,000 dollars	NIL	https://tab.gl adl y.io
6.	Buoy Up	Signup, open news -> wherever there is red highlight-click & donate, vetted charities picked by experts (charity navigator, give well, GuideStar, open philanthropy project)	Bbc, Cnn, Forbes etc.	NIL (user signup, website - TBR)	https://buoy up. org/

The team referred to various resources for literature review of the product as well as past research perspectives. A few of the existing products followed the same application, but none of the products matched the exact application.

The entire literature survey is divided into two major sections. Section [A] focuses on the pre-existing products available in the market and their comparison with respect to specifications, features, customers using the product, and cost. Although these products do not provide the function that our recommendation engine provides, many pre-seed ideas mentioned depict the in-line cause of either donation or selected charity recommendation partly.

[1] WPForm-

WP Forms is a drag-and-drop form builder that is commonly added to an existing website and is used by over 4 million websites. It's a customizable form that includes one-time donations as well as signup forms, email marketing services, and other features.

[2] GiveWP-

GiveWP is a one-time donation platform with Google integrations and fine-tuned features such as customizable donation tributes, such as "in honour of." The most basic paid package is $240 per year.

[3] Paypal-

Paypal is a WordPress plugin that allows you to link your PayPal account to your WordPress site. It supports a variety of payment methods, including recurring monthly donations.

[4] Woocommerce-

Woocommerce focuses on many levels of contributions and allows you to select the lowest and maximum payment levels, designate a fixed amount for donation, or let consumers specify their own amount, as well as place the choice on the product, cart, or checkout pages, or utilise a widget.

[5] Tab for a cause-

The platform includes a fantastic feature that allows you to donate to a charity each time you open a new browser tab. Woocommerce got 3,000 users and gave $4,000 dollars, according to their stats.

[6] Buoyup –

Whenever a user wakes up and checks the news, they may click on the red highlight on the item to donate to vetted organizations chosen by professionals, such as charity navigator, give well, and Guidestar open philanthropy.

B) Journals and Research papers

The research articles that were referred to during the study and implementation are the topics of the second portion of the literature review.

[1] The first citation - 15th International Joint Conference on Computer Vision, Imaging, and Computer Graphics Theory and Applications (VISIGRAPP '20): **Building a Data Visualization Recommender System** has defined, specified, and gone through the process of creating a data visualisation recommender system in Valletta, Malta. The authors meticulously researched current solutions, performed a survey, compiled specifications for a model, built a model, tested it, assessed it, and finally created an example implementation. They've demonstrated that data visualisation recommender systems have a role in the data science industry. This article acts as a step-by-step instruction manual for someone who wants to make a data visualisation recommender system.

[2] "Usage of Visualization Techniques in Data Science Workflows (Data Management Technologies and Applications)" pushes for improved communication between the two domains of data science and visualisation research. Users can benefit significantly from visual interfaces that align with information. However, there is still a "Gap in Interactive Visualization" for exploratory data analysis.

This was also demonstrated by the study on the use of visualisation approaches in standard data science tools described in this publication. Interviews with data scientists, on the other hand, suggest a strong desire to use innovative methodologies to get fresh insights into their datasets. As a result, a variety of approaches are proposed for better integrating visualisation techniques into popular data science procedures.

[3] The research article Naive Bayes classifier approach is utilised for classification in our third reference. **"Recommendation System Using Naive Bayes Classifier."** The Naive Bayes technique, which is a probabilistic strategy, is used to examine each record in the dataset, and the results are then given in the form of a list. The recommendation method is based on good and negative remarks that have been categorised. The classifier is very scalable and just requires a few parameters. The Naive Bayes Classifier is used to compute the overall number of positive and negative reviews in a given review collection. The proposed method might be used to assess sentiment in any type of review dataset. It is also possible to extend the job to a higher degree of the input set.

[4] "A new approach to context-based recommendations using support vector machine techniques" The purpose of this research is to apply methods that can be effectively integrated into the system, based on both contextual and non-contextual information. Is to show you how to do it. (Gujrat Technological University, Gujrat, India, Faculty of Computer Science and Engineering, 2015). In this task, we used the SVM classification model to combine context-dependent and context-independent user settings.

[5] The following citation, "A Comparative Analysis of Machine Learning Algorithms for Recommender Systems" proposes pub mender, a publishing recommender system. The accuracy of this approach was shown to be 329 percent greater than that of Journal Finder.

[6] The next reference - **Recommender Systems: An Overview, Research Trends, and Future Directions** presents a thorough examination of the RS, including several recommendation methodologies, related difficulties, and information retrieval strategies. Because of its vast applicability, it has piqued the curiosity of a large number of scholars all over the world. The primary goal of this report is to identify RS research trends. From 2011 through the first quarter of 2017, almost 1000 research articles published by ACM, IEEE, Springer, and Elsevier were evaluated. This study yielded a number of intriguing insights that will aid present and future RS researchers in evaluating and planning their research agendas. This work also looks ahead to the future of RS, which might lead to new research areas in this field.

[7] Source of New Ideas for Charity Fundraising: Empirical Research Rejuvenating Lost Donors and First Donors Due to Fierce Competition in the UK Charity Fundraising Market and the Almost Constant Needs for Charity to Attract New Supporters Convince them to get additional donors Donations, charities are forced to start new fundraising activities on a regular basis. A combination of concepts or ideas that attracts public attention, differentiates charities from other funding organizations, and differentiates charities' current marketing efforts from the past is the starting point for innovative campaigns. The question of where and how funding organizations get ideas for new campaigns is a big issue in the area of charity idea generation, but it's often overlooked.

[8] Bela Gipp, Joeram Beel and Christian Hentschel have presented detailed study on Scienstien. Scienstein is the **First Hybrid Research Paper on Recommender System** and a potent alternative to presently used academic search engines, as described in this paper.

Scienstein improves on the commonly used keywordbased search strategy by including citation analysis, author analysis, source analysis, implicit and explicit ratings, as well as novel and yettobeused methodologies like as the `DistanceSimilarity Index' (DSI) and the 'Intext Impact Factor' (ItIF). Instead of merely inputting keywords, a user may upload full papers, including reference lists, and rate them implicitly or explicitly to enhance suggestions. Similar and related papers may readily be determined using citation, author, and source analysis. A user-friendly GUI manages all these approaches.

[9] Another reference, Aman kumar, Sharma Assistant, and Suruchi Sahni's **Comparative Study of Classification Algorithms for Email Data Analysis**, explains that email has become one of the most efficient and cost-effective means of communication in recent years. I am. However, due to the increase in email users, spam emails have increased significantly in recent years. Emails are classified as spam or non-spam using a data mining classification algorithm. In this study, we used the WEKA environment to test four algorithms for spam email datasets (ID3, J48, Simple CART, and Alternating Decision Tree). Next, we compared the four methods in terms of classification accuracy. Simulation results show that the J48 classifier is superior to ID3, CART, and ADTree in terms of classification accuracy.

[10] The empirical study of Irina Rish's Naive Bayes Classifier has the overall goal of understanding the data characteristics that affect naive Bayes performance. The given method uses Monte Carlo simulation. This allows you to thoroughly investigate the classification accuracy of various randomly generated problems. The effect of distribution entropy on classification errors is investigated and shown that feature distributions with low entropy generate high naive Bayesian power. Naive Bayes also works well for near-feature feature dependencies and performs best in two situations: completely independent features (as expected) and feature-dependent features (surprisingly). It also shows that it does. Another unexpected result is that the accuracy of naive Bayes is independent of the degree of characteristic dependency evaluated by the class-related mutual information between the characteristics.

3. Specifications

3.1 Language- Python

3.2 IDE & Other Softwares used- Google Colaboratory

- GitHub

- Jira Software

3.3 Libraries used- pandas, matplotlib, NumPy, missingno, Altair, streamlit

3.4 Datasets- Scraping the Charity Navigator webpages which have information of recognized charities and creating a dataset(csv file) from the scraped information.

4. Methodology

4.1 Real Life View of the system

Figure 4.1(a) Real Life View of the System

Figure 4.1(b) Real Life View of theSystem

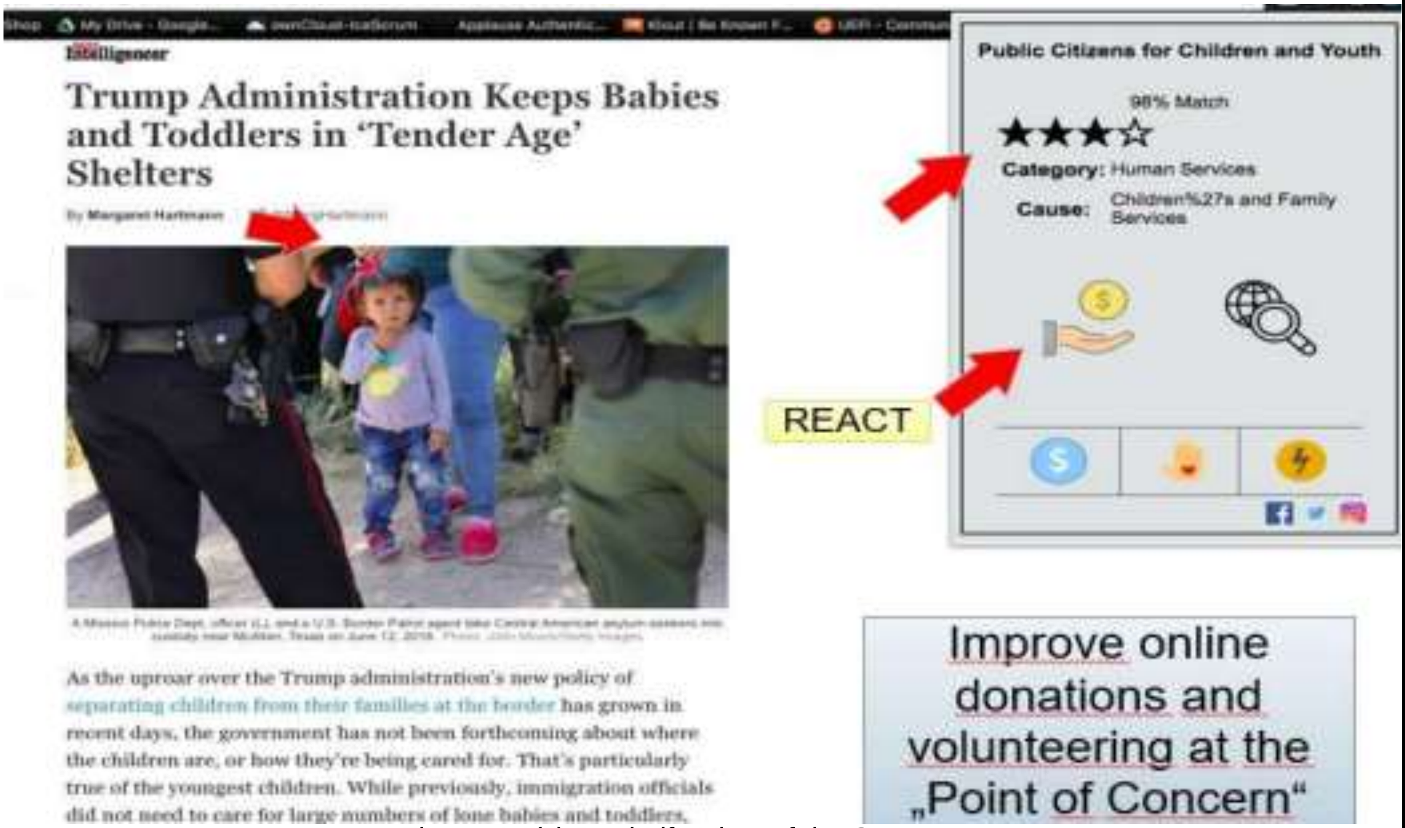


Figure 4.1(c) Real Life View of the System

Figure 4.1(a), (b) and (c) represent the real-life view of the model, once the user integrates the plugin with any web browser, the plugin will analyze the article present on the user's screen and a pop-up box showing the donation, similar article recommendations, and other services will pop up, the symbols in bottom show payment and support options with user's total credits in the first section of the plugin.

Every time the user opens a new article, information will pass through the backend model and the plugin will process the information and recommend new articles and services in an optimized manner.

4.2 Block Diagram of the system

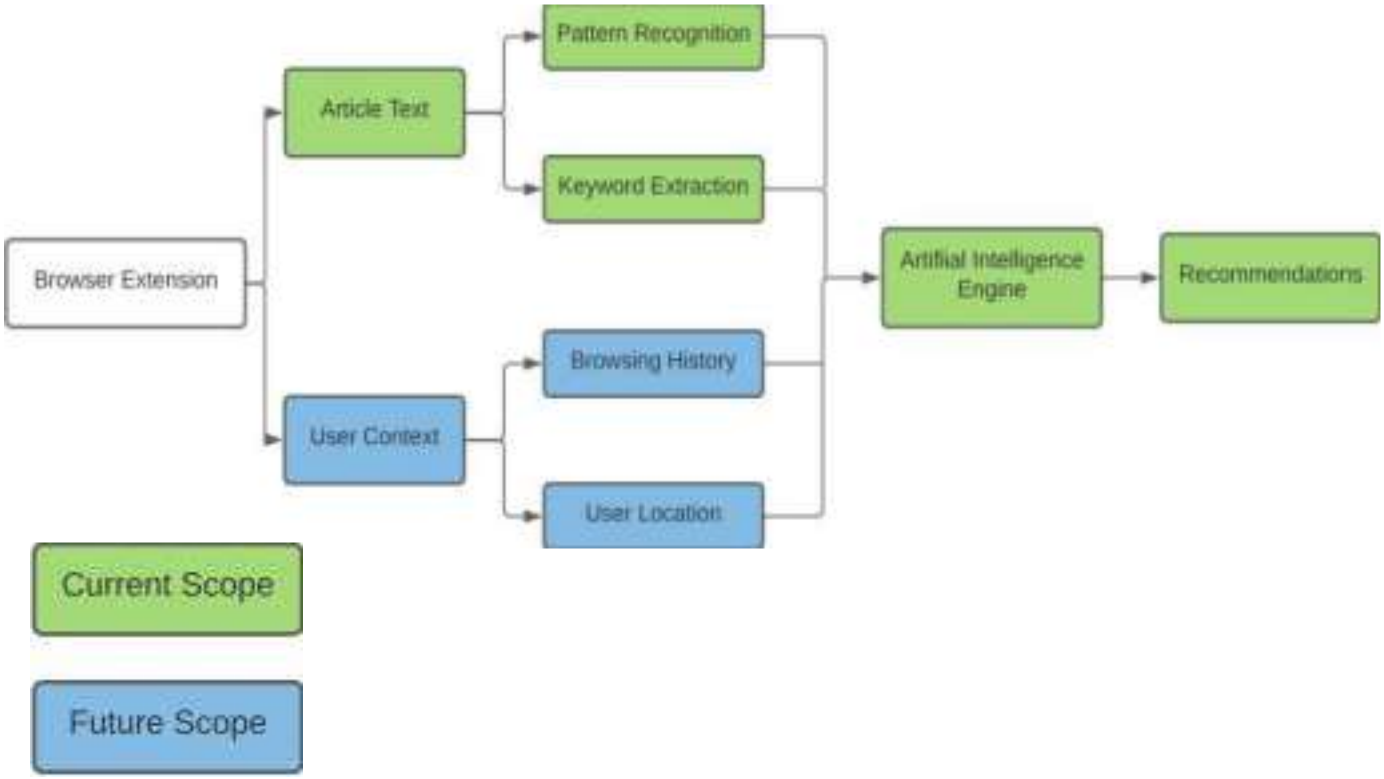


Figure 4.2 Block Diagram of the System

The above block diagram represents the detailed flow of the technical implementation of the system with the current and future scope that the product will follow. Article sentiment analysis, pattern recognition, keyword extraction, and recommendation system are the primary goals of the project.

Detailed Explanation of the Block Diagram:

1. **Article Text:** The text from the integrated article will be extracted and analysed
2. **Pattern Recognition:** To understand the sentiment and correlation between the same category articles, pattern recognition will take place.
3. **Keyword extraction:** To find out the relation between causes and article, a bag of words will be created which will help in matching the training and testing data information.

4. **Recommendation Engine:** Based on the keywords and article sentiment cause category an AI recommendation engine supporting the flow of article recommendations and charity donations will be built, the focus is on providing real-time value to the user.
5. **User Context:** Browsing history and user location are considered under the future scope of the model to provide highly personalized recommendations to the user.

4.3 Working of the system

These are the steps explaining the working of the system:

1) The input for the model will be text from online news articles.
2) We will be using APIs from trusted sources like Charity Navigator to get a list of numerous verified charities for several social causes. We will be scraping the data from web pages that have information regarding recognized charities such as - Charity name, Website link, Ratings, Category/Cause supported, Donation link, and other relevant information. Further, we will be creating a dataset(CSV file) out of the scraped information.
3) NLP and pre-processing techniques such as
    a) Visualising and analysing patterns in the data
    b) Checking class balance
    c) Dropping unwanted rows/ columns
    d) Stemming words and filtering integers

   will be applied to the contents of the dataset.
4) This processed data will be fed to a machine learning model (classifier) that will predict the social cause that the article talks about. The classifiers will be trained on the category/ cause.
5) Further the accuracy of the model will be calculated along with the

   classification report which contains different performance metrics such as precision, recall, and f1 score that are calculated from the confusion matrix.

6) The model will then be tested on Text from random Web-Articles.
7) Lastly, Depending on the predicted cause top 3 charities will be recommended to the user based on the ratings.
8) The model can be further integrated with a Web-browser Plugin or a Web-Application.

4.4 Integration of Model with Interface

[A] Grafana Plugin

Moving ahead with the additional objective of a user-friendly dashboard, our development included the plugin creation. The team explored various tools and languages which included Grafana platform integration. Grafana supports a wide range of usage logs including plugins, traffic metrics, and other cloud applications. Connecting the plugin through this platform would lead our data science model recommendation system to take the inputs and pop the user board of location, donation, and recommend desired blocks to the user.

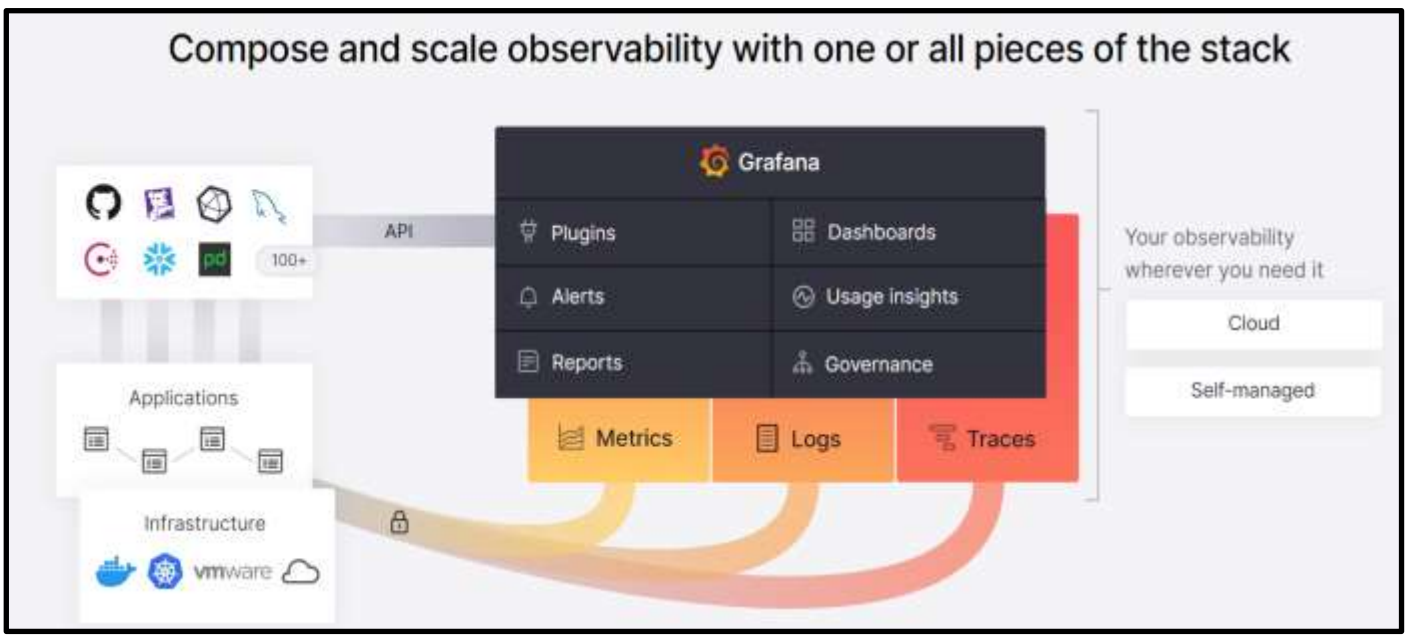


Figure 4.3 Grafana Dashboard (src: https://grafana.com/)

Grafana labs support a wide range of data sources like MySQL, Datadog, etc. and the required skills include data source implementation in the front end, Grafana 7.0, Go 1.14+, Mage, NodeJS, yarn, etc. Since the main focus of the project included developing a classification and recommendation engine which pondered on the data science skills, the time limitations, resource scarcity, and platform constraints led the team to re-evaluate the plugin development part leading to a better alternative for tunnel website application development.

[B] Web-Application (Tunnel Website)

A software that runs on a web browser is a Web Application, unlike a regular application that runs locally on the user system. Web applications are preferred when storage is a concern. We will be developing a Web-App (tunnel website) using ***Streamlit.***

5.Detail Design

5.1 Software Design: - Algorithm

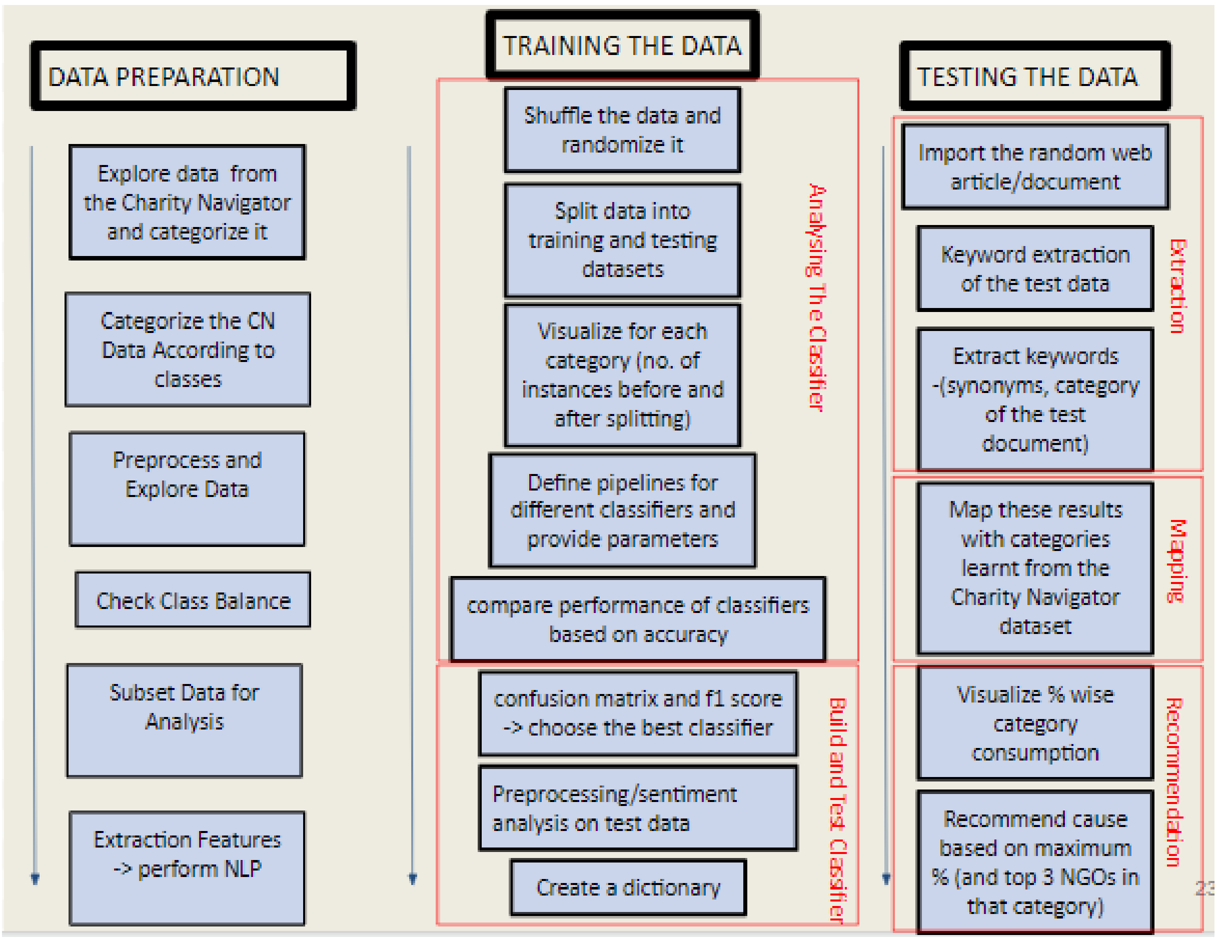


Figure 5.1 Algorithm / Workflow

Following is the detailed explanation of the Workflow:

1. Data Preparation: First step of the detailed design is exploring the data and categorizing it according to the classes, in this step we are pre-processing and exploring the data to perform sentiment analysis and analyses the subset class balance of every category according to the charity navigator dataset. The features and then extracted using NLP and processes for sentiment analysis take place in the data cleaning and preparation phase.

2. The training stage: The data is shuffled and randomized for training the classification algorithms, pipelines have been defined with well-settled parameters and the performance is checked on confusion and f1 matrix for five algorithms explained in the later stage.

3. Testing The Data: The entire model is then tested on the algorithm, and keywords from the article are extracted and mapped with the results from trained data. Further, we will be visualizing every category based on the percentage of the cause and the final result will be recommended as the next article for users with donation options. Algorithms used for comparative analysis are:

The classifiers are used to train the model and predict the social cause highlighted in the Web- Article. We have used and compared 5 different classifiers on our dataset.

Classifiers:

A) Naive Bayes

It is a learning algorithm that is supervised. The naive Bayes classifier is used as a stochastic classifier. It is based on a probabilistic model that incorporates strong independence assumptions.

It is a family of algorithms that share a common principle: each pair of features to be categorised is independent of the others. The basic premise is that each feature contributes equally and independently to the output/final result
It is based on the Bayes Rule or Bayes law, which is used to calculate the likelihood of a hypothesis based on prior knowledge and conditional probability.

P(A/B) = (P(B/A) * P(A)) / P(B)

[P(A)] is Prior Probability (Hypothesis before observing the evidence.)
[P(B)] is Marginal Probability (Evidence.)

For Naïve Bayes, the model predicted an accuracy of 0.95 in our case.

B) Logistic Regression

Logistic regression is a predictive analytics method based on the concept of probability. This is a machine learning algorithm used for classification problems. The 21 sigmoid function is used in logistic regression. It is used to model the data.

h = 1 / (1 + pow(e,-1))

This method typically produces a number between 0 and 1. A 0.5 threshold can be used. If the sigmoid function produces a number that is more than or equal to 0.5, we consider it to be 1, and if it returns a value that is less than 0.5, we consider it to be 0.

A multinomial probability distribution is the probability distribution that defines multi-class probabilities. Through the model, we were able to obtain an accuracy of 0.96 using Logistic Regression.

C) Support Vector Machines

SVM stands for Supervised Machine Learning and can be used to solve classification or regression problems. It transforms the data using a technique called kernel tricks and calculates the ideal boundaries between possible outputs based on those changes.

SVMs work by using a (nonlinear) mapping function to transfer data in input space to data in feature space, allowing a problem to be linearly separated. The SVM then finds the best separating hyperplane (which can be a complex decision surface when projected back into input space via 1). SVMs are intriguing since they have a solid theoretical foundation as well as cutting-edge success in real-world applications.

For SVM, our model predicted an accuracy of 0.97.

D) Stochastic Gradient Descent

The iterative gradient descent method is an algorithm. Start at a random location on the feature and work downhill until you reach the bottom of the feature. This algorithm is useful when you cannot find the optimal position just by making the slope of the function equal to 0.

The main concept is to start with a random point (in our parabola example, a random "x") and figure out how to update it with each iteration so that we descend the slope.The algorithm's accuracy was 0.97 according to the model.

E) Random Forest

Random forest is a supervised learning technique that can be used to classify and predict data. It is based on ensemble learning.

This is a manner to combine a big quantity of classifiers to resolve complicated troubles and enhance version performance. Random forests are classifiers that integrate a fixed of choice timber from extraordinary subsets of a dataset and common the consequences to enhance the predictability of the dataset. Random woodland collects predictions from every tree rather than counting on a unmarried choice tree. Predict the very last output primarily based totally on the bulk of the votes cast. The more the quantity of timber withinside the woodland, the extra correct it'll be and the avoidance of overfitting troubles. The model received a 0.95 f1-score and 0.84 accuracy for the random forest algorithm.

4. In order to improve the accuracy of the models following steps are implemented:

a) Training the model on the category instead of the cause

Initially the models(classifiers) were trained on the causes. However, the causes that were synonymous with each other were grouped into a single main category to improve the accuracy of the models.

b) Improving the Class Imbalance

A lot of class imbalance was observed between the different categories that were present in the dataset-

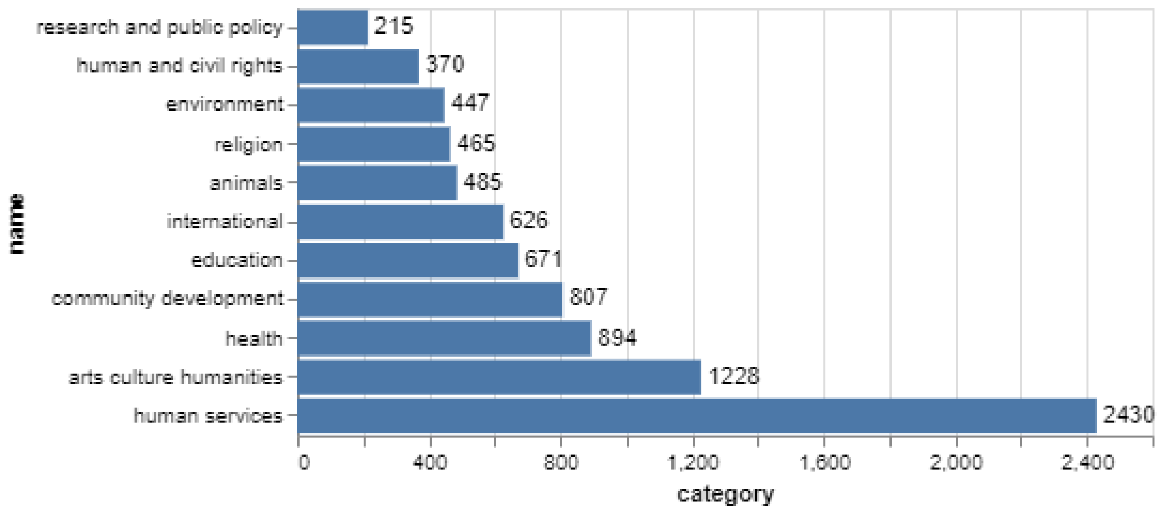


Figure 5.2 Class-Imbalance

To improve this class imbalance we tried applying **SMOTE** or **Synthetic Minority Oversampling Technique.**

SMOTE follows a simple approach of duplicating examples in the minority class by the process of over-sampling. However, this technique failed to improve the class imbalance as our models were dealing with text data. SMOTE appears to be problematic on text data because it works in a feature space and its output is not synthetic data which is a real representative of a text inside its feature space.

As an alternative to SMOTE, we resolved the issue of class imbalance using the method of sampling. We chose a threshold value and under-sampled the classes having data points larger than this threshold value and over-sampled the classes having data points lower than this threshold value to achieve balanced classes.

As a result, the accuracy of all the models increased significantly.

5. Recommending Charities

Charities that support the predicted cause are recommended based on the following metrics:

a) Overall Rating

b) Financial Rating

c) Rating for Accountability and Transparency Based on these metrics the charities of the given cause are sorted. As a result, Top 3 charities in the 3 sorted lists are recommended.

6. Integrating with Web app

Streamlit is an Open-Source Python framework used to develop ML and Data Science Web-Applications. One of the advantages of using this framework is that the code is written in Python, much easier than learning a different stack.

The Web-App reruns whenever the code is modified. The framework has widgets available by default. Streamlit builds great apps with less code since it has many utilities already available. We built a sample web-app to learn how streamlit works. This web-app plots the closing price and volume of Google stocks from 2011 to 2020.The web-app/website is accessible as long as the Google Colab notebook is running. Following are the snippets of the Sample Stocks Web-Apps:

URL generated by the Sample web-app:

```
!streamlit run myapp.py & npx localtunnel --port 8501

2022-05-18 16:47:25.111 INFO    numexpr.utils: NumExpr defaulting to 2 threads.

  You can now view your Streamlit app in your browser.

  Network URL: http://172.28.0.2:8501
  External URL: http://35.236.164.67:8501

npx: installed 22 in 5.539s
your url is: https://silver-mails-show-35-236-164-67.loca.lt
```

Figure 5.3 Sample Web-app URL

Images of the Sample Stocks Web-App:

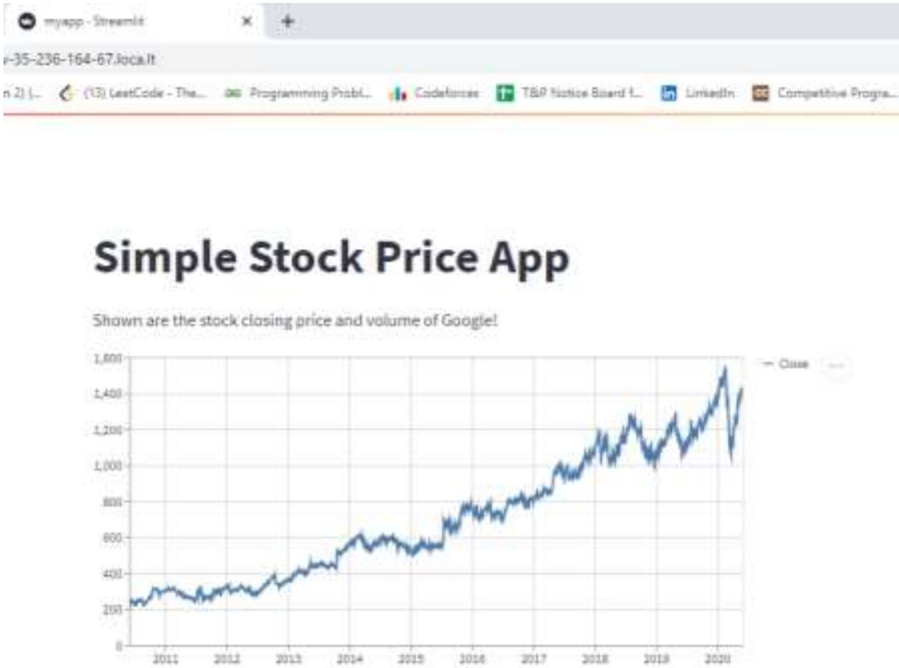


Figure 5.4(a) Sample Stocks Web-app

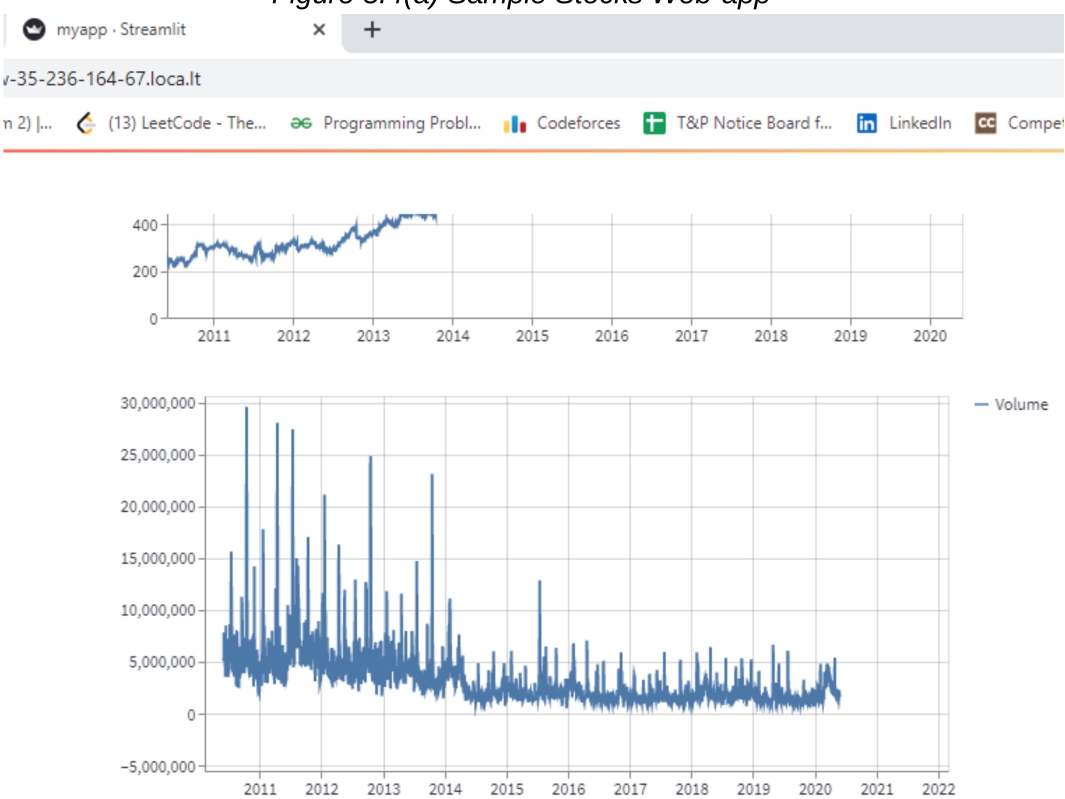


Figure 5.4(b) Sample Stocks Web-app

We integrated the developed Recommendation Engine with a Web-App

6. Results

1) Visualizing Missing Rows

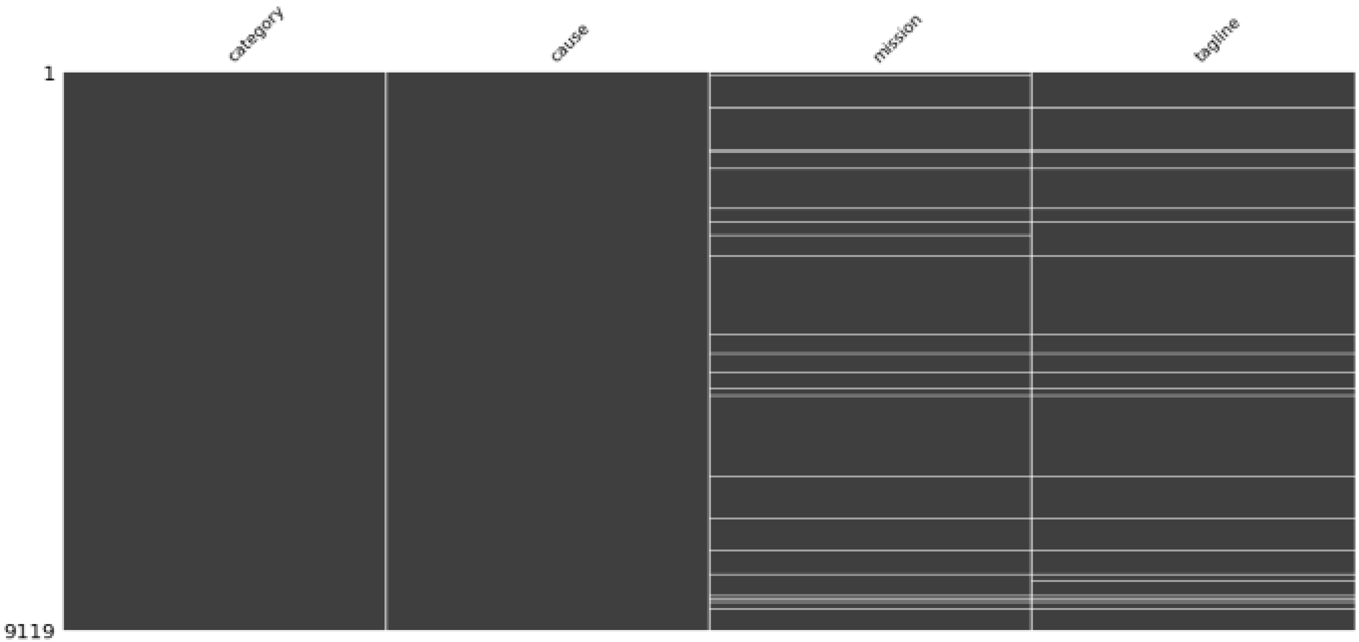


Figure 6.1 Missing Rows

The figure above represents the missing rows in the dataset. A line is shown in the figure to represent the data points for which the mission and tagline are missing. There are no missing values for the category and cause column for all the data points in the dataset.

2) Dropping Missing Rows-

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8638 entries, 0 to 9118
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   category  8638 non-null   object
 1   cause     8638 non-null   object
 2   mission   8638 non-null   object
 3   tagline   8638 non-null   object
dtypes: object(4)
memory usage: 337.4+ KB
```

Figure 6.2 Dropped Missing Rows

Missing rows have been dropped from the dataset to balance it throughout.

3) Checking Class Balance

	category	name
human services	2430	human services
arts culture humanities	1228	arts culture humanities
health	894	health
community development	807	community development
education	671	education
international	626	international
animals	485	animals
religion	465	religion
environment	447	environment
human and civil rights	370	human and civil rights
research and public policy	215	research and public policy

Figure 6.3 Checking Class Balance

This table shows the number of data points/ records that belong to a particular category and the figure below shows the graphical representation of the same.

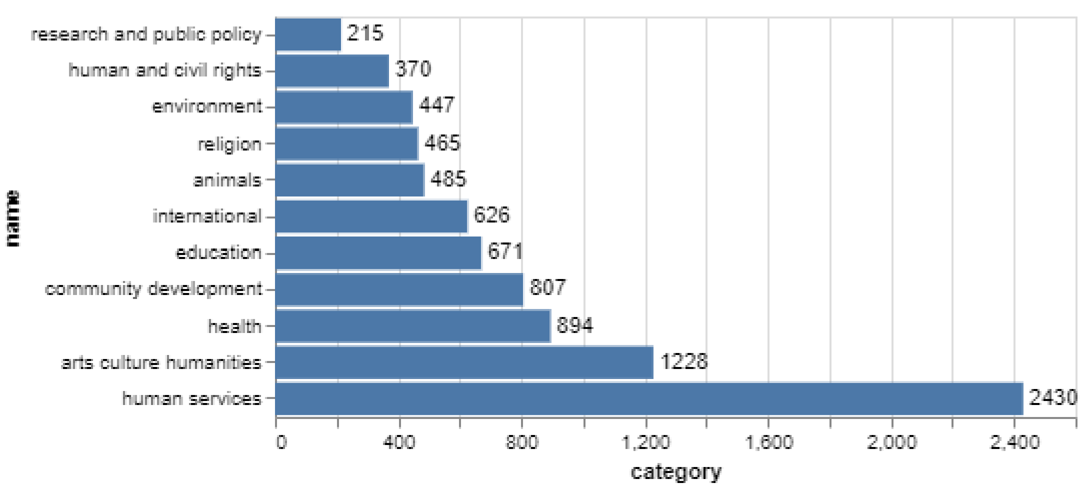


Figure 6.4 Social Causes

4) Comparing Classifiers

We have first compared classifiers based on accuracy. Since our dataset is small, accuracy can be deceiving. Hence, we have further compared their performance based on confusion matrices and F1 scores. An N x N matrix is used to assess the overall performance of a type version, in which N is the quantity of goal classes. The matrix compares the real intention values to the gadget mastering version's predictions. This affords us with a complete photograph of the way properly our type version is running and the styles of mistakes it makes.

Following are the results of the confusion matrix for each of the algorithms, along with the F1 scores. It should ideally have 0 in all non-diagonal positions and the diagonal values should be high. However, we have some non-zero values at the non-diagonal positions in the confusion matrix of each classifier. Precision and Recall are the 2 constructing blocks of the F1 rating. The intention of the F1 rating is to mix the precision and don't forget metrics right into a unmarried metric. At the identical time, the F1 rating has been designed to paintings properly on imbalanced data. In our model, Naive Bayes and Random Forest show reliable score and accuracy. Hence, the model is being tested on the same.

A) Naive Bayes

```
Test accuracy is 0.9528795811518325

 CONFUSION MATRIX
[[55  3  1  1  0]
 [ 0 70  9  0  1]
 [ 2  0 87  0  1]
 [ 0  0  0 79  0]
 [ 0  0  0  0 73]]

CLASSIFICATION REPORT
              precision    recall  f1-score   support

           0       0.96      0.92      0.94        60
           1       0.96      0.88      0.92        80
           2       0.90      0.97      0.93        90
           3       0.99      1.00      0.99        79
           4       0.97      1.00      0.99        73

    accuracy                           0.95       382
   macro avg       0.96      0.95      0.95       382
weighted avg       0.95      0.95      0.95       382
```

The precision accuracy for the Naive Bayes Classifier is 0.953. It has a confusion matrix that shows high diagonal values with a good F1 score. Some non-diagonal values in the confusion matrix are low but not zero.

B) Logistic Regression

```
Test accuracy is 0.9581151832460733

CONFUSION MATRIX
[[59  0  1  0  0]
 [ 0 68 12  0  0]
 [ 0  2 87  0  1]
 [ 0  0  0 79  0]
 [ 0  0  0  0 73]]

CLASSIFICATION REPORT
              precision    recall  f1-score   support

           0       1.00      0.98      0.99        60
           1       0.97      0.85      0.91        80
           2       0.87      0.97      0.92        90
           3       1.00      1.00      1.00        79
           4       0.99      1.00      0.99        73

    accuracy                           0.96       382
   macro avg       0.97      0.96      0.96       382
weighted avg       0.96      0.96      0.96       382
```

The prediction accuracy for Logistic Regression is 0.958. The non-diagonal values are not all zero, diagonal values in confusion matrix are lower than those of Naive Bayes but the F1 score is better.

C) Support Vector Machines

```
Test accuracy is 0.9659685863874345

CONFUSION MATRIX
[[60  0  0  0  0]
 [ 0 71  9  0  0]
 [ 0  3 86  0  1]
 [ 0  0  0 79  0]
 [ 0  0  0  0 73]]

CLASSIFICATION REPORT
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        60
           1       0.96      0.89      0.92        80
           2       0.91      0.96      0.93        90
           3       1.00      1.00      1.00        79
           4       0.99      1.00      0.99        73

    accuracy                           0.97       382
   macro avg       0.97      0.97      0.97       382
weighted avg       0.97      0.97      0.97       382
```

The prediction accuracy for Support Vector Machines is 0.967 with a good confusion matrix and F1 score.

C) Stochastic Gradient Descent

```
Test accuracy is 0.9633507853403142

CONFUSION MATRIX
[[59  1  0  0  0]
 [ 0 71  8  1  0]
 [ 0  3 86  0  1]
 [ 0  0  0 79  0]
 [ 0  0  0  0 73]]

CLASSIFICATION REPORT
              precision    recall  f1-score   support

           0       1.00      0.98      0.99        60
           1       0.95      0.89      0.92        80
           2       0.91      0.96      0.93        90
           3       0.99      1.00      0.99        79
           4       0.99      1.00      0.99        73

    accuracy                           0.96       382
   macro avg       0.97      0.97      0.97       382
weighted avg       0.96      0.96      0.96       382
```

The accuracy for Stochastic Gradient Descent is 0.971 with a good F1 score and confusion matrix values.

E) Random Forest

```
Test accuracy is 0.7905759162303665

CONFUSION MATRIX
[[59  0  0  0  1]
 [ 8 35 15  4 18]
 [ 5  3 56  5 21]
 [ 0  0  0 79  0]
 [ 0  0  0  0 73]]

CLASSIFICATION REPORT
              precision    recall  f1-score   support

           0       0.82      0.98      0.89        60
           1       0.92      0.44      0.59        80
           2       0.79      0.62      0.70        90
           3       0.90      1.00      0.95        79
           4       0.65      1.00      0.78        73

    accuracy                           0.79       382
   macro avg       0.81      0.81      0.78       382
weighted avg       0.82      0.79      0.77       382
```

The accuracy for Random Forest is 0.838. The model has a good confusion matrix and F1 score.

Based on the above results, we have concluded that **Naive Bayes and Random Forest are the best-suited classifiers for our data** since they have high accuracies, high diagonal values for confusion matrices as well as good F1 scores.

5) Balanced Classes achieved after under-sampling and over-sampling:

```
animals                          671
arts culture humanities          671
community development            671
education                        671
environment                      671
health                           671
human and civil rights           671
human services                   671
international                    671
religion                         671
research and public policy       671
Name: category, dtype: int64
```

Figure 6.5 Balanced Classes

This figure represents the balanced classes that are achieved after under-sampling and over-sampling. Each class now has an equal number of data points(671).

6) Updated values of Performance Metrics After Sampling:

A) Naive Bayes

```
Test accuracy is 0.96274217585693

 CONFUSION MATRIX
[[138   0   0   0   3]
 [  1 114   0   1   4]
 [  0   0 132   5   0]
 [  1   5   3 124   1]
 [  0   1   0   0 138]]

CLASSIFICATION REPORT
              precision    recall  f1-score   support

           0       0.99      0.98      0.98       141
           1       0.95      0.95      0.95       120
           2       0.98      0.96      0.97       137
           3       0.95      0.93      0.94       134
           4       0.95      0.99      0.97       139

    accuracy                           0.96       671
   macro avg       0.96      0.96      0.96       671
weighted avg       0.96      0.96      0.96       671
```

The accuracy for Naive Bayes has increased from 95.2 to 96.2 and the f1-score has also increased successfully.

B) Logistic Regression

Test accuracy is 0.977645305514158

CONFUSION MATRIX
```
[[140   0   0   0   1]
 [  1 116   0   1   2]
 [  0   1 134   2   0]
 [  0   5   1 127   1]
 [  0   0   0   0 139]]
```

CLASSIFICATION REPORT
```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       141
           1       0.95      0.97      0.96       120
           2       0.99      0.98      0.99       137
           3       0.98      0.95      0.96       134
           4       0.97      1.00      0.99       139

    accuracy                           0.98       671
   macro avg       0.98      0.98      0.98       671
weighted avg       0.98      0.98      0.98       671
```

After achieving balanced classes, the confusion matrix for logistic regression has improved along with a significant improvement in the accuracy and f1-score.

C) Support Vector Machines

Test accuracy is 0.9836065573770492

CONFUSION MATRIX
```
[[140   0   0   0   1]
 [  1 117   0   1   1]
 [  0   0 136   1   0]
 [  0   4   0 129   1]
 [  1   0   0   0 138]]
```

CLASSIFICATION REPORT
```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       141
           1       0.97      0.97      0.97       120
           2       1.00      0.99      1.00       137
           3       0.98      0.96      0.97       134
           4       0.98      0.99      0.99       139

    accuracy                           0.98       671
   macro avg       0.98      0.98      0.98       671
weighted avg       0.98      0.98      0.98       671
```

The accuracy has increased by 2% for SVM.

D) Stochastic Gradient Descent

```
Test accuracy is 0.9806259314456036

CONFUSION MATRIX
[[140   0   0   0   1]
 [  1 116   0   2   1]
 [  0   0 135   2   0]
 [  0   4   0 129   1]
 [  1   0   0   0 138]]

CLASSIFICATION REPORT
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       141
           1       0.97      0.97      0.97       120
           2       1.00      0.99      0.99       137
           3       0.97      0.96      0.97       134
           4       0.98      0.99      0.99       139

    accuracy                           0.98       671
   macro avg       0.98      0.98      0.98       671
weighted avg       0.98      0.98      0.98       671
```

SGD also showed a significant improvement in all the performance metrics and high values at the diagonal(True Positives) of the confusion matrix which shows that the model has improved at predicting the cause correctly.

E) Random Forest

```
Test accuracy is 0.8882265275707899

CONFUSION MATRIX
[[108   4   0   2  27]
 [  0 116   0   3   1]
 [  2   3 124   8   0]
 [  1  14   5 112   2]
 [  0   2   1   0 136]]

CLASSIFICATION REPORT
              precision    recall  f1-score   support

           0       0.97      0.77      0.86       141
           1       0.83      0.97      0.90       120
           2       0.95      0.91      0.93       137
           3       0.90      0.84      0.86       134
           4       0.82      0.98      0.89       139

    accuracy                           0.89       671
   macro avg       0.90      0.89      0.89       671
weighted avg       0.90      0.89      0.89       671
```

Random Forest showed significant improvement in the accuracy from 83.8 to 88.8% and it also shows a good confusion matrix.

7) Comparison Of Accuracies after Sampling:

Table 6.1 Comparison of Accuracies

Classifier	Accuracy Before Sampling	Accuracy After Sampling
1) Naive Bayes	95%	96%
2) Logistic Regression	96%	97.7%
3) Stochastic Gradient Descent	97%	98.2%
4) Support Vector	96%	98.3%
5) Random Forest	81%	88.8%

The table above shows the comparison of accuracies of each model before and after sampling. Each model shows a significant increase in accuracies.

8) Testing on Unknown Data-

```
[ ]  museum_str = "Museums are a peek into the history of a country, and in our case
     children_str = "Save the Children is India’s leading independent child rights’ N

     education_str = "Even after 65 years of independence, Millions of Indian childre
     education_str_edited = "even after 65 years of independence millions of indian c
     og_education_str = "life issues institute is dedicated to changing hearts and mi

     environment_str = "the student conservation association is the only national org

     animals_str = "The world's last male white rhino is almost certain never to mate

     arts_cult_hum_str = "Historic preservation plays a key role in building vibrant
```

Figure 6.6 Testing on Unknown Data

Article text is stored as a string in different variables which is given as an input to the classifiers.

Output of the models when tested on unknown data-

```
results = SGD_pipeline.predict([children_str, education_str, education_str_edited, og_education_str, museum_str, environment_str, animals_str, arts_cult_hum_str])
for i in results:
  print(target_to_category[i])

education
education
education
environment
arts culture humanities
environment
animals
arts culture humanities
```

Figure 6.7 Checking Model Output

The figure shows the output of the models when tested on unknown data. Figure 6.7 shows the output of SGD and how accurately it predicts when tested on unknown data.

9) Integration with Web app

a) Generated URL:

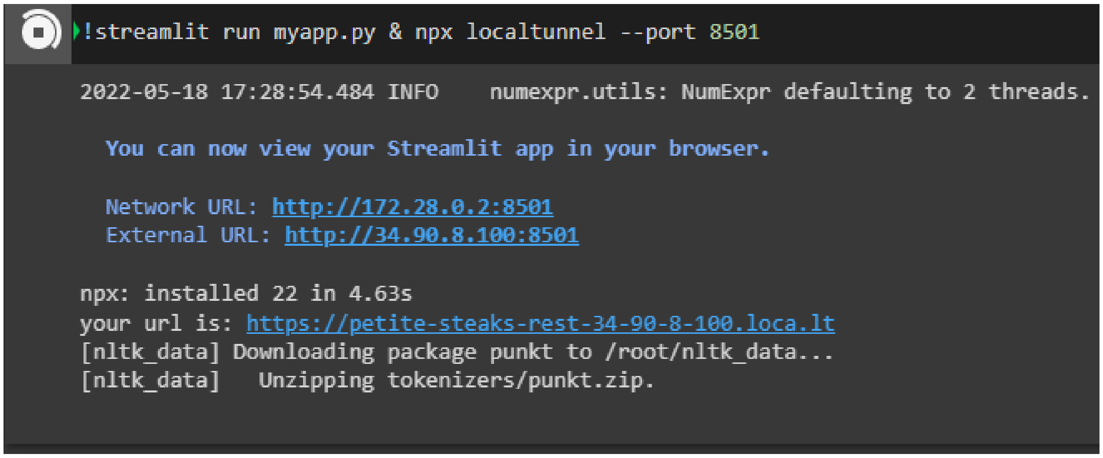


Figure 6.8(a) Web-App URL

After the execution of the code, the web app generates a URL which is a tunnel website.

b) Enter the article link

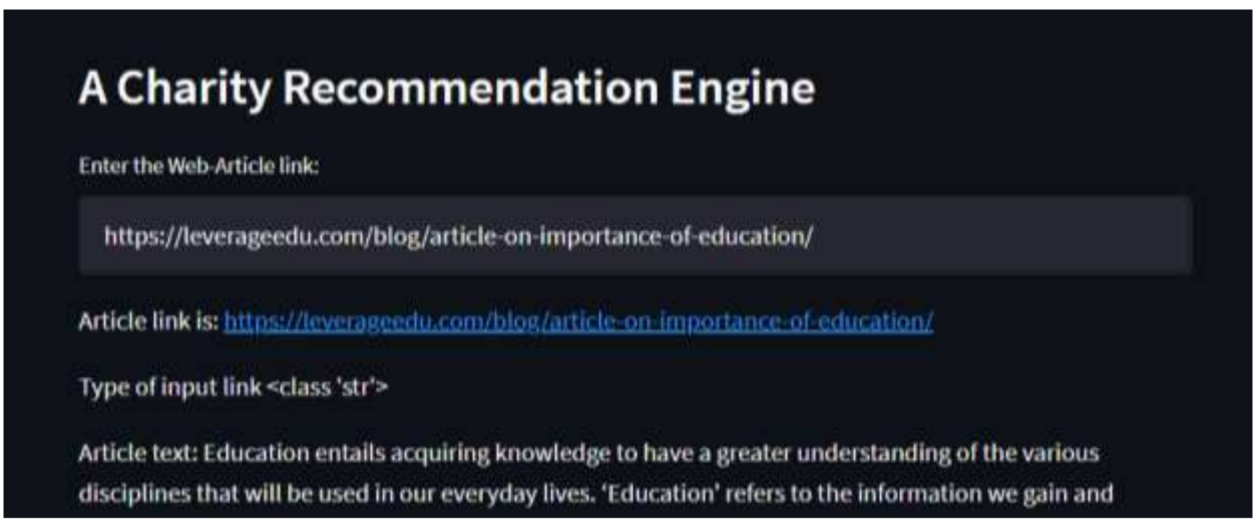
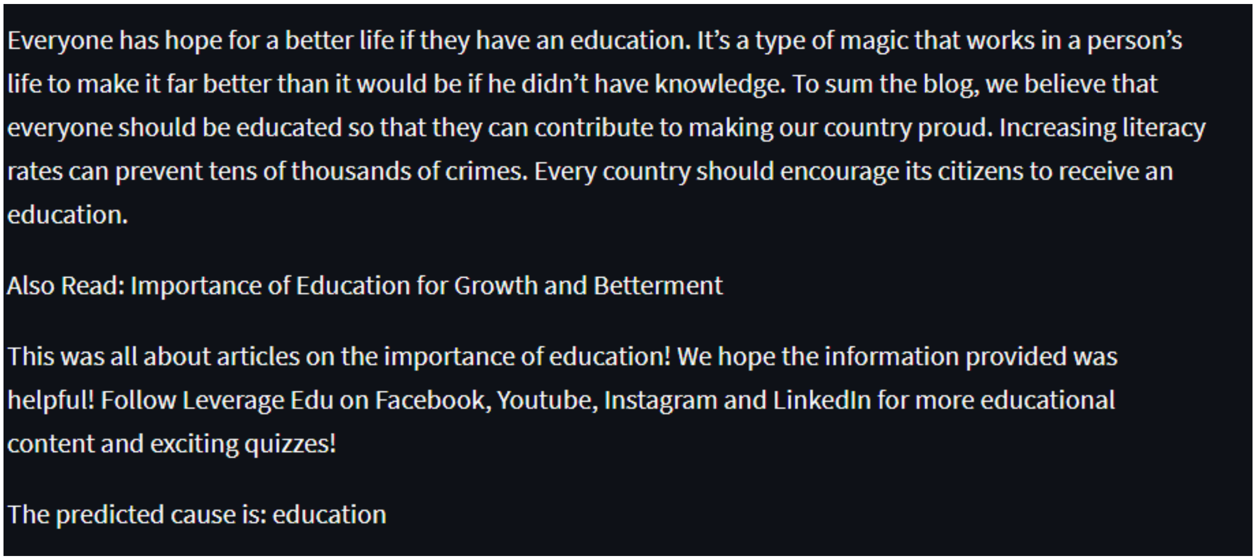


Figure 6.8(b) Charity Recommendation Web-App (Input)

After clicking on the generated URL the web app asks to enter the article link. After entering the article link result will be generated.

c) Predicted Cause


Everyone has hope for a better life if they have an education. It's a type of magic that works in a person's life to make it far better than it would be if he didn't have knowledge. To sum the blog, we believe that everyone should be educated so that they can contribute to making our country proud. Increasing literacy rates can prevent tens of thousands of crimes. Every country should encourage its citizens to receive an education.

Also Read: Importance of Education for Growth and Betterment

This was all about articles on the importance of education! We hope the information provided was helpful! Follow Leverage Edu on Facebook, Youtube, Instagram and LinkedIn for more educational content and exciting quizzes!

The predicted cause is: education

Figure 6.8(c) Charity Recommendation Web-App (Prediction)

The figure above shows the result generated by the web app - the predicted cause and the article text.

d) Recommended Charities

Figure 6.8(d) Charity Recommendation Web-App (Recommendation)

The 3 figures above show the final result - Recommended Charities.

Top 3 charities are recommended that support the cause of the article depending on the

1. Overall Rating
2. Financial Rating
3. Accountability and Transparency Rating

e) Recommended Charity

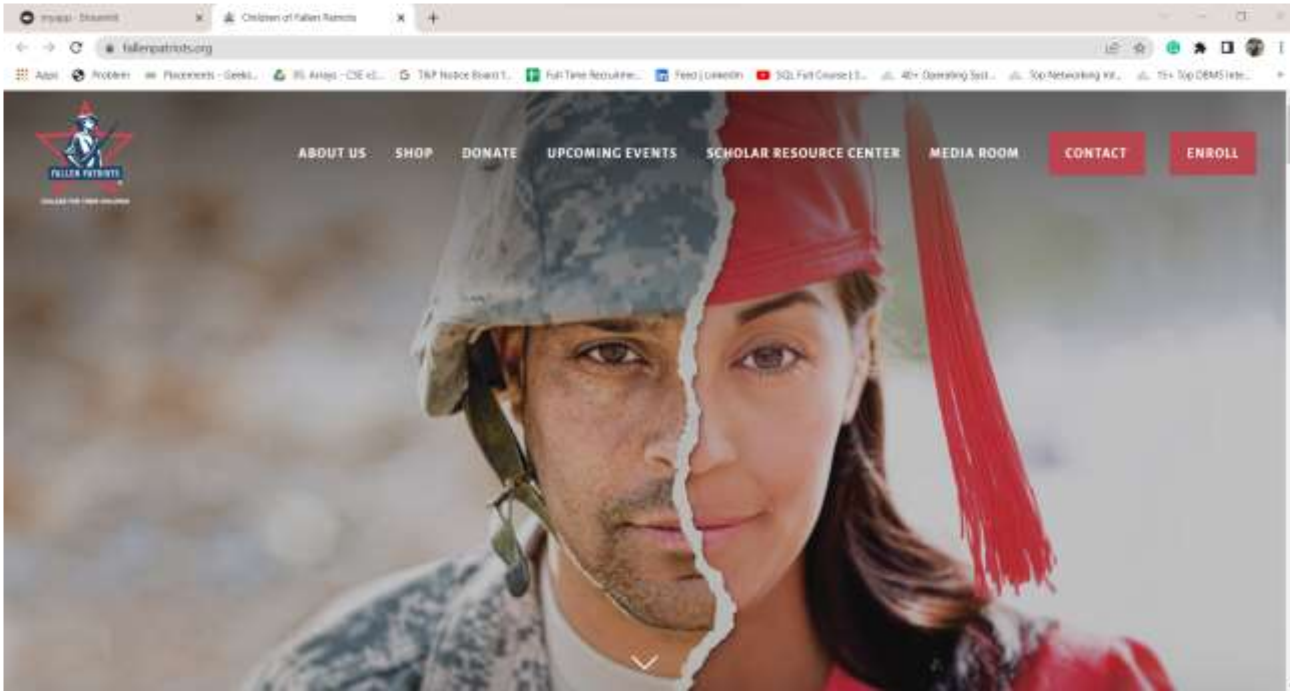


Figure 6.8(e) Recommended Charity

f) Donation Page

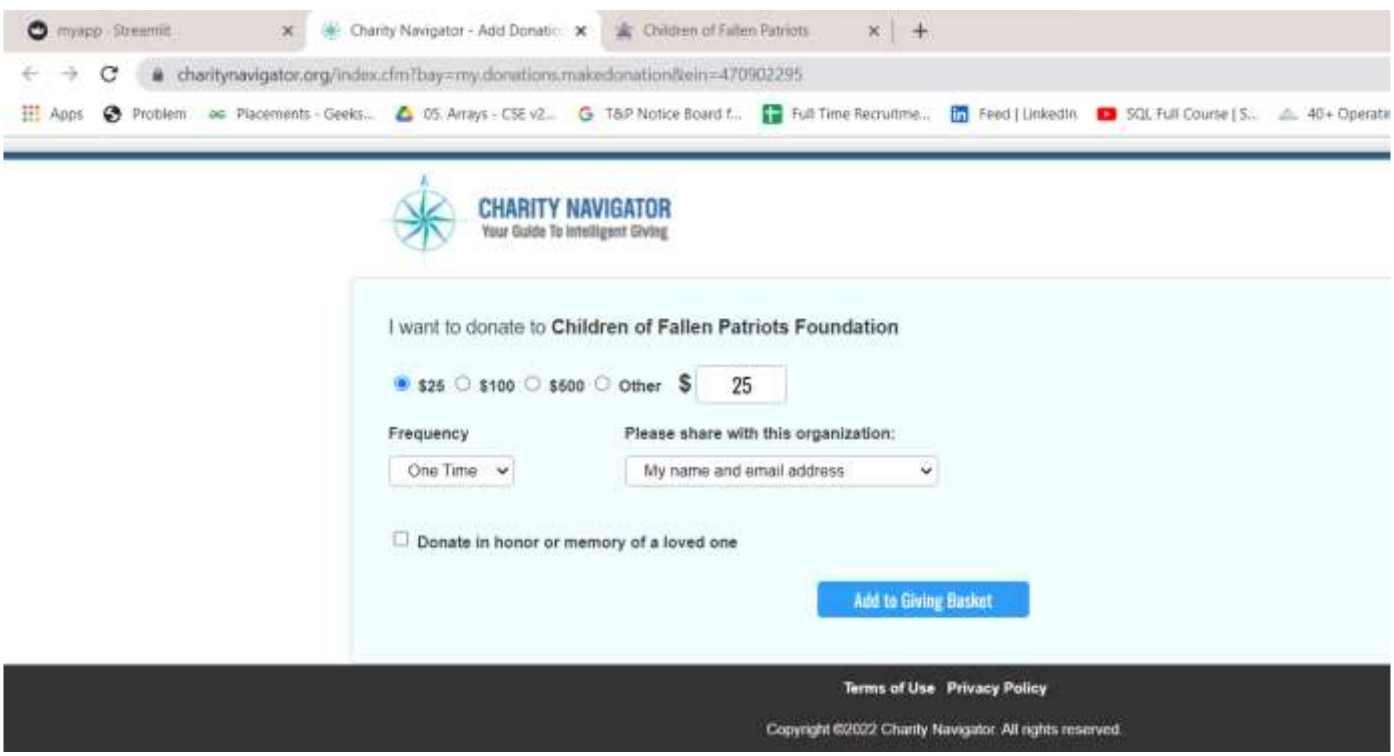


Figure 6.8(f) Charity Donation Page

7. Limitations

1. The available dataset contains a limited number of data points/ records which is one of the major fallbacks in training, our model has been trained on limited data points. However, we will expand the dataset as well as generalize the categories and causes in order to get a more accurate prediction accuracy.

2. We have trained our model only on 5 classes/ categories and not on the entire dataset so that we can have a subset of data for analysis.

3. Another limitation is that we have to paste the article link which is given as an input to the web app and then the model predicts the output and recommends the charities compromising the user experience.

8. Conclusion

Our aim was to develop a model that recommends relevant and recognized charities based on the web articles read by a user online. We have tried and tested five classifiers viz. Naive Bayes classifier, Random Forest Classifier, SVM classifier, SGD Classifier, and Logistic Regression Classifier for training as well as testing on unknown data.

Based on the prediction accuracy, confusion matrix, and F1 score of each, we have inferred that Random Forest and Naive Bayes are the two best classifiers for our case. Further, work is being done on the testing part.

A thorough comparison of plugin and web application development was performed and after the significant improvement in testing unknown data, the team has integrated the model with the tunnel website application.

9. Future Scope

1. Customization-

Customization to users can be provided based on their preferences, locations, etc. This will not only give better results but will also help local charities in expanding their reach in neighbouring regions.

2. Integrating the AI model with a Browser extension/App-

This will provide a better User Experience and make donating to users' choice of charity easier and more convenient. Developing an Android Application or a Web Plugin for the model will minimize effort for the user, thus maximizing the number of net donations made.

3. User Experience-

This can be improved by modifying the Web-App / Browser Plugin. Better visuals as well as User Interface can be provided.

4. **Recommendations based on previous donations**-

This can be provided to increase number of donations and help more NGOs supporting similar social causes.

10. Bibliography

1. “How are Logistic Regression & Ordinary Least Squares Regression (Linear Regression) Related? Why the “Regression” in Logistic?” https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

2. Mohammed Seyam, Faridah Othman, Ahmed El-Shafie , “Prediction of Stream Flow in Humid Tropical Rivers by Support Vector Machines" presented at the MATEC Web of Conferences, January 2017, 111(10):01007.

3. SGD “Gradient Descent and Stochastic Gradient Descent Algorithms for Neural Networks

4. ‘https://medium.com/konvergen/gradient-descent-and-sstochastic-gradient-descent-algorithms-for-neural-networks-e817f3c411ef

5. Daniel James Mennit, Kirk Sherril, Kurt M Fristrup, “ A geospatial model of ambient sound pressure levels in the contiguous United States’ presented in the Journal of Acoustical Society of America, May 2014, 135(5):2746-2764.

6. Petra Kubernátová, Magda Friedjung ová, Max van Duijn., constructing a Data Visualization Recommender System (15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP ’20) At: Valletta, Malta, Volume: Volume 3: IVAPP), July 2019.

7. Valletta, Malta. Usage of Visualization Techniques in Data Science Workflows (Data Management Technologies and Applications), Conference: 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP ’20), January 2020.

8. Mr. Pandian C, Dept. of Information Technology, K.L.N. College of Engineering, Tamil Nadu, India, Recommendation System Using Naive Bayes Classifier, March 2020, IRJET-V7I31110.

9. Satya Prakash Sahu, Anand Nautiyal, Mahendra Prasad, Machine Learning Algorithms for Recommender System- A comparative Analysis, International Journal of Computer Applications Technology and Research 6(2):97-100, August 2021.

10. Aansi A. Kothari, Warish Patel, A Novel Approach Towards Context Based Recomendations Using Support Vector Machine Methodology, (Department of Computer Science and Engineering, Gujrat Technological University, Gujrat, India, 2015), Procedia Computer Science 57:1171-1178, December 2015.

11. Satya Prakash Sahu, Anand Nautiyal, Mahendra Prasad, Machine Learning Algorithms for Recomender System - A comparative Analysis, Presented in “International Journal of Computer Applications Technology and Research, 10.7753/IJCATR0602.1005, February 2017.

11. Work Plan

a)Work plan

Task / Module — Month & Week no.	Sign. of Int. Guide	Sign. of the Co-ordinator	Sept 3rd week	Oct 1^{st} week	Oct 2^{nd}, 3^{rd} week	Nov 1^{st}, 2^{nd} week	Dec 1^{st}, 2^{nd} week	Jan 1st week	Feb 1^{st} week	Feb 2^{nd}, 3rd week	March 1^{st} week	March 2^{nd}, 3^{rd} week	April 1^{st} week	April 2^{nd} week	April 3^{rd} week	May 1st week	May 2nd week	May 3^{rd} week	June 1^{st} week
1. Literature Survey & Concept development ☐			■	■															
2. Specifications of ENTIRE SYSTEM ☐					■	■													
3. System-level design (Block Schematic) ☐						■													
4. Real-life Layout of the System ☐																			
5. Hardware Components Selection & Procurement ☐																			
6. Detail design (Hardware Design / Software Programming of each Block) ☐							■	■	■										
a) Break up Task 6 ☐																			
b) into suitable stages																			
7. Validation of each module's Design (Hardware & or Software) thro'. SIMULATION ☐											■	■							
8. Seminar Report (with Plagiarism Check) ☐																			
9. Internal Seminar ☐																■			
10. P.C.B. design (for Hardware Part) ☐																			
11. Integration & Testing of the Complete System ☐																			
12. Internal Project Demo. ☐																			
13. Project Report (with Plagiarism Check) ☐																		■	

Figure 11.1 Workplan

12. Plagiarism Report

turnitin

Similarity Report ID: oid:8054:17900272

PAPER NAME

Group 52-Final Project Report .docx

WORD COUNT

7976 Words

CHARACTER COUNT

42158 Characters

PAGE COUNT

62 Pages

FILE SIZE

4.3MB

SUBMISSION DATE

May 30, 2022 3:57 PM GMT+5:30

REPORT DATE

May 30, 2022 3:58 PM GMT+5:30

9% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 5% Internet database
- 1% Publications database
- Crossref database
- Crossref Posted Content database
- 8% Submitted Works database

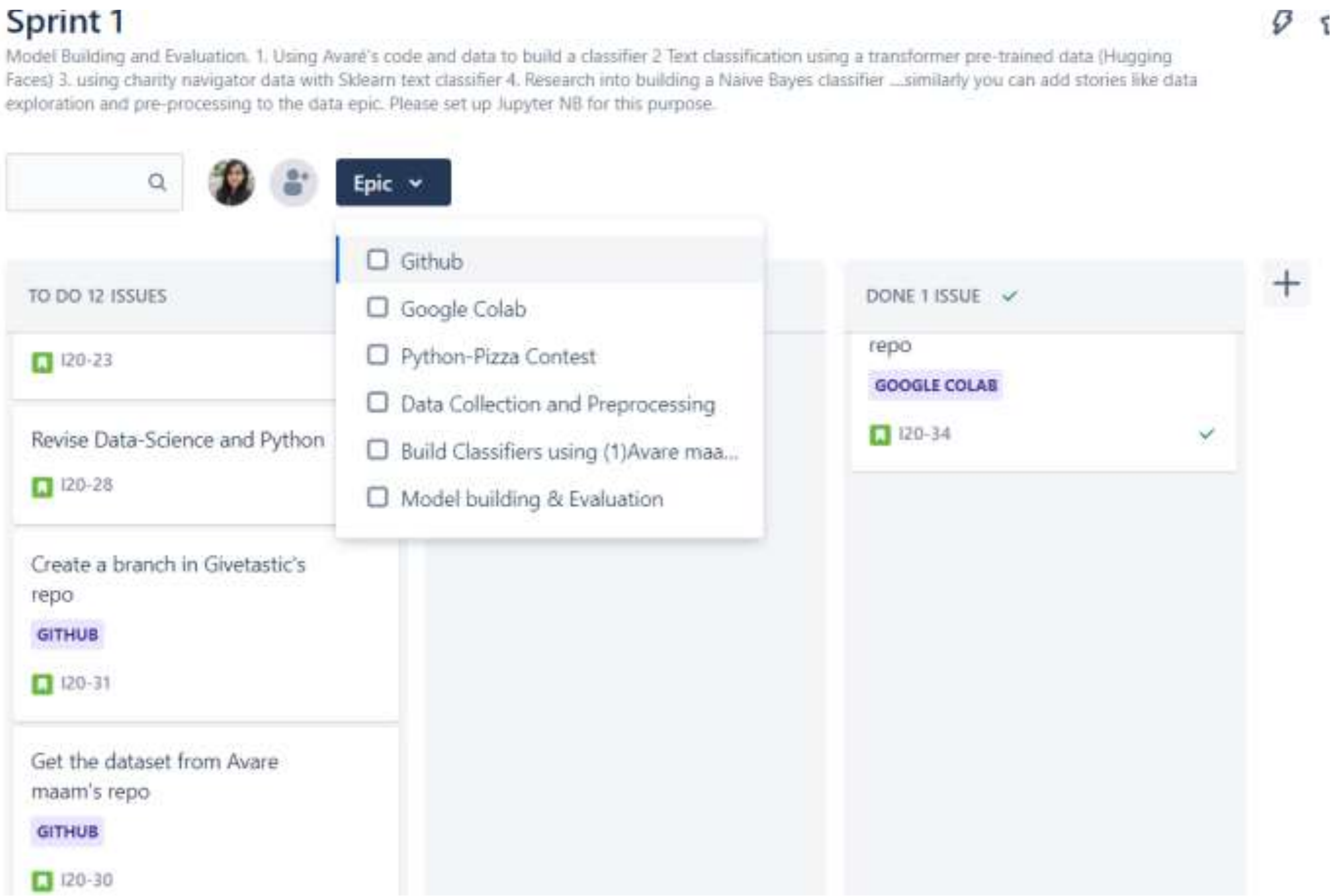
Excluded from Similarity Report

- Manually excluded sources
- Manually excluded text blocks

Figure 12.1 Plagiarism Report

Appendix I

Agile Management using Jira Software



Agile Development- Jira Board

Agile is a methodology which focuses on work management in a sophisticated and smart approach leading to a structured approach in product development. A scrum master defines the scope and role of project members and allows dedicated sprints for workflow. Every sprint is a dedicated goal to be covered by the team involving technical and non-technical objectives. The closing goal is split into numerous stages and controlled in agile technique which ends up in steady collaboration with stakeholders and non-stop development over the period. Agile technique makes a speciality of four centre standards: people and interactions over tactics and tools, operating software program over complete documentation, consumer collaboration over agreement negotiation, and responding to alternate over following a plan. Following this process, the team was successfully able to perform incremental improvement through smalls and frequent releases. Following image shows the sprint plan. The team has used Jira software for sprint planning, task distribution, and software integration.

Appendix II

Python Pizza, Hamburg: Micro-conference-

The idea was presented in a remote conference on the 11th December, 2021 where the team presented the product methodology and ideation. Presenters and attendees from various countries admired the idea with constructive feedback and improvement suggestions.

Python Pizza Micro-Conference

Appendix III

Frankfurt Investors (Product Pitch)

The team pitched the project as a product with respect to its technology as well as customer recognition side on the 10th of February, 2022 and received valuable inputs from mentors and investors. A good insight into the financial perspective was advised.

Pitching for Frankfurt Investors

Appendix IV

Poster Presentation

MKSSS's Cummins College of Engineering for Women, Pune. - The project structure was presented at the College level on the 22nd of February, 2022, where the team described the workflow, methodology, and entire structure of the project along with information about two micro conference presentations, agile methodology, and future perspective of the project.

Poster Presentation