**PROJECT REPORT**

**ST. XAVIER'S COLLEGE [AUTONOMOUS]**

**KOLKATA**

**TEXT CLASSIFICATION USING A.I. TECHNIQUES**

**Debabrata Datta[1] , Aniket Patra[2] , Arnanta Chatterjee[3] , Sirshendu Ganguly[4]**

**[1] Assistant Professor, Department of Computer Science, St. Xavier's College, Kolkata (India).**

**[2] PG Scholar, Roll No: 520, [3] PG Scholar, Roll No: 545, [4] PG Scholar, Roll No: 525**

#### Submitted to

#### The Department of Computer Science

#### in partial fulfillment of

#### the requirements for the degree of M.Sc

## CERTIFICATE OF AUTHENTICATED WORK

This is to certify that the project report entitled “Text Classification using A.I. Techniques” submitted to Department of Computer Science , ST. XAVIER'S COLLEGE [AUTONOMOUS], KOLKATA, in partial fulfillment of the requirement for the award of the degree of MASTER OF SCIENCE (M.SC.) is an original work carried out by Mr Aniket Patra, Arnanta Chatterjee, Sirshendu Ganguly, Registration no. 012-1121-3889-15, 563-1121-2141-15, 544-1121-0083-15 under my guidance. The matter embodied in this project is authentic and is genuine work done by the student and has not been submitted whether to this College or to any other Institute for the fulfillment of the requirement of any course of study.

#### 

#### Signature of the Student 1: Signature of the Professor

………………………………… ………………………………………….

#### Signature of the Student 2: Date: ……………….

#### 

#### …………………….

#### Signature of the Student 3:

#### 

#### …………………… Name and Designation of Professor

#### Date:………………….. ………………………………………………… ..………………………………………………. ……..………………………………………….

**ROLES AND RESPONSIBILITIES FORM**

**Name of the Project** “Text Classification Using A.I Techniques”

**Date:**

|  |  |  |
| --- | --- | --- |
| **Name of the Team Member** | **Role** | **Tasks and Responsibilities** |
| 1. Aniket Patra | Data and hardware manager | Arranges different sources of different data for the project and manages the emails back and forth. |
| 1. Arnanta Chatterjee | Team co-ordinator and auditor | To administer the work and assign task to the team members which also includes myself and made sure to complete given objectives on the deadlines assigned. |
| 1. Sirshendu Ganguly | Design head and auditor | To understand the objective of the project and design the models accordingly. |

Name and Signature of the **Project Team members:**

#### 1. Name: ……………………………….. Signature………………..

#### 2 Name: ………………………………. Signature…………………

#### 

#### 3. Name: ………………………………. Signature …………………

#### Signature of the Professor:……………………… Date: …………………..

**ABSTRACT**

The task of labeling books according to their genre is a challenging classification problem, having in mind that genre is an immaterial feature that cannot be directly pinpointed in any of the book summary. Moreover , books may belong to multiple genres at the same time at the same time, making book genre assignment typically a multi-label classification problem, which is much more tedious than standard single-label classification. In our project we first set our multi-label classification problem statement and then normalize the dataset by implementing stemming at the initial stage as it would allow the classifier to give a better accuracy as it would allow the classifier to treat different forms of the same word as the same concept and then use a strategy to build a book genre prediction model and then we create an inference function so that when new book summaries come in the future, our prediction system should be able to take a book summary as input and generate appropriate genre tags.

**ACKNOWLEDGEMENT**

The project could not have been successfully completed without the wholehearted support of our guide Prof. Debabrata Dutta. He always boosted our morals assiduously. We thank sir for this support and encouragement during the development of the project. He is a mentor who is very helpful. We admire his infinite patience and understanding that he guided us in a field we had no previous experience about.

We would also like to thank other professors and faculty members of Computer Science Department who have rendered us significant encouragement. We thank them all for the guidance and other staffs who directly and indirectly assisted us in the successful completion of our project “**Text classification using A.I. Techniques**”.

**TABLE OF CONTENTS**

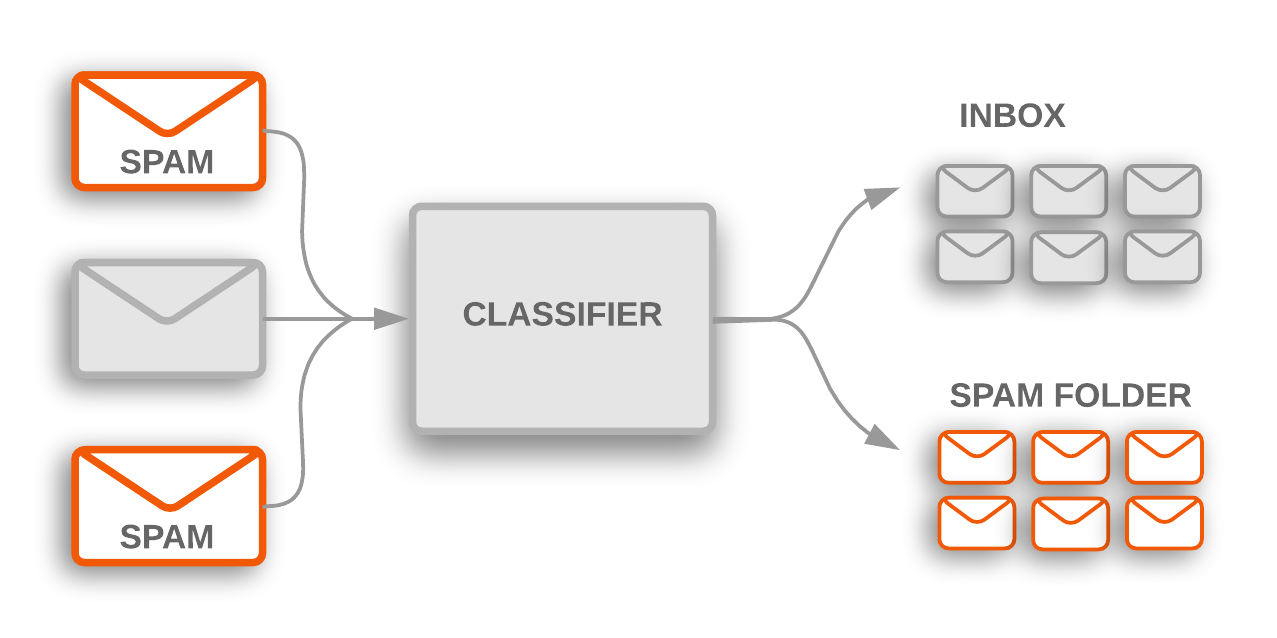
|  |  |
| --- | --- |
| **Content** | **Page No:** |
| **CHAPTER 1: INTRODUCTION** | **8 - 11** |
| **CHAPTER 2: SURVEY OF TECHNOLOGIES** | **12 - 13** |
| **CHAPTER 3: REQUIREMENTS AND ANALYSIS** | **14 - 21** |
| **CHAPTER 4: SYSTEM DESIGN** | **22 - 25** |
| **CONCLUSION** | **26** |
| **REFERENCES** | **27 - 28** |

**TABLE OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Description** | **Page No.** |
| Figure 1 | Text classification filtering spamming | 8 |
| Figure 2 | Survey on choice of Programming language | 12 |
| Figure 3 | NumPy vs Pandas performance | 13 |
| Figure 4 | NumPy and its dependencies | 13 |
| Figure 5 | Project planning and scheduling steps | 16 |
| Figure 6 | Relation between time and phases of a project | 18 |
| Figure 7 | Gantt chart | 19 |
| Figure 8 | Conceptual Model of our project | 21 |
| Figure 9 | Flowchart of our model | 24 |

**CHAPTER 1: INTRODUCTION**

At the cutting edge of computer science, Natural Language Processing (NLP) has become indispensable, as it gives machines the capability to understand the text, which was so far the exclusive privilege of humans. Text Classification also known as Text Categorization is the process of classifying textual information by means of its content. It is actually the task of sorting a set of documents into classes from a predefined dataset. This task falls at the intersection of information retrieval and machine learning techniques, and has witnessed a thriving interest from researchers and data scientists. Due to the humongous nature of data, text classification finds its application in various domains such as information extraction , text summarization, information retrieval , sentimental analysis and many more. We encounter unstructured data almost everywhere such as social media, emails, web pages, survey reports and many more. Documents containing such texts are going to use different forms of a single word, mainly because to abide grammar of that particular language in which the document is written. In order to achieve higher precision, it becomes necessary to reduce such words to a common base form. This process is termed as stemming.



**Figure 1: Text Classification is used to flag incoming spam mails, which are filtered into a spam folder [1].**

Text Classification can be broadly classified into three main categories. These are Unsupervised text classification, Supervised text classification and Semi-supervised text classification. In Unsupervised text classification, no information is provided externally whatsoever, whereas in Supervised text classification information related to the correct text classification is provided by some mechanism which is external to the classification model . Thus it becomes easier to test the accuracy of the text classification model. In Semi-supervised text classification parts of the text are labeled by some external mechanism.

There are mainly two factors which make text classification a demanding task :- (I) feature extraction and (II) topic ambiguity. Firstly , the feature extraction mainly deals with taking out the most appropriate set of features that accurately describe the document and helps to build a good classification model. Secondly, many big documents are so complicated that it becomes tedious to put it into any specific category. Also, in broad documents there maybe certain terms that have different meanings based on their usage in different contexts and occurs multiple times within the document in different contexts. **[1]**

**Background:** Over the past decades, volume of data has been exponentially increasing. The need to allow researches to extract insight from the data promoted the implementation of mechanism to support the task. In **[2]**, H.P Luhn, stated that in order to determine which sentences of an article serves as the auto-abstract, a measure is required by which the information content of all the sentences can be graded and compared. Today, these mechanism are known as information retrieval processes.

In **[3]** , a binary text classifier had been implemented, where the newly classified instances of data and their predicted classes were added to the training data set so the enhanced data set would be able to train the classifier for future unclassified instances of data. This allows the classifier to increase the accuracy each time the classifier was used for similar problems. The primary limitation of classifier used in **[3]**  is that it can perform binary classification only, but in the real world a classifier needs to be dynamic and a classifier should not constrain the data set to any specific number of classes .This limitation was removed in **[4]**, where a multi-class classifier was used along with tf-idf statistic **[5][6].**

An addition that can be added to the classifier is to incorporate stemming into the training and classification processes. This would allow the classifier to give a better accuracy as it would allow the classifier to treat different forms of the same word as the same concept.In many cases the dataset is not one dimensional, in those cases we have to use the concepts of multi-label classification. The apparent concept is in the name itself – multi-label classification, here an instance can have multiple labels and the number is not fixed unlike binary classifier where the number of classes is two, or multi-class classifier where there are more than two labels but only one label is there for each input.

In the tables provided below, ‘X’ represents the input variable and ‘Y’ represents the target variable.

**Table 1: Binary Classification**

|  |  |
| --- | --- |
| **X** | **Y** |
| X1 | t1 |
| X2 | t2 |
| X3 | t1 |
| X4 | t2 |
| X5 | t1 |

In Table 1, since there are only two labels t1 and t2, here ‘Y’ is a binary target variable, and it is an example of binary classification.

**Table 2: Multi-class Classification**

|  |  |
| --- | --- |
| **X** | **Y** |
| X1 | t2 |
| X2 | t3 |
| X3 | t4 |
| X4 | t1 |
| X5 | t3 |

In Table 2, since there are more than two labels but only one label is there for each input, it is a multi-class classification.

**Table 3: Multi-label Classification**

|  |  |
| --- | --- |
| **X** | **Y** |
| X1 | [t1,t5] |
| X2 | [t1,t2,t3,t4] |
| X3 | [t3] |
| X4 | [t2,t4] |
| X5 | [t1,t3,t4] |

In Table 3, there are multiple labels not only across the table but for individual value of ‘X’, thus it is a multi-class classification.

**Objective :** The goal of the project is used to determine the genre of a book from its summary by using multi-label text classification. Books are not one dimensional , one book can span several genres. We have extracted book summary and got to work using the concepts of multi-label classification.

**Purpose, Scope and Applicability:**

* + **Purpose:-**  In this era, Natural Language Processing NLP has become indispensable at the hand of data scientists, as it gives machines the capability to understand the text, which was so far the exclusive privilege of humans. We have used Text Classification along with NLP to predict book genre accurately using book summary dataset.
  + **Description:-** Binary classifiers can only predict , when the target variable has two labels only, but we know that books are categorized into more than two genre, thus we had to improve on the previous work using the concepts of multi-label classification.
  + **Scope:-** In our project we first set our multi-label classification problem statement and then normalize the dataset by implementing stemming at the initial stage as it would allow the classifier to give a better accuracy as it would allow the classifier to treat different forms of the same word as the same concept and then use a strategy to build a book genre prediction model and then we create an inference function so that when new book summaries come in the future, our prediction system should be able to take a book summary as input and generate appropriate genre tags.
  + **Applicability:-** This project will be able to predict book summaries that come in the future.

**Achievements:** Explain what knowledge you achieved after the completion of your work. What contributions has your project made to the chosen area? Goals achieved - describes the degree to which the findings support the original objectives laid out by the project. The goals may be partially or fully achieved, or exceeded.

**Organisation of Report:** Summarising the remaining chapters of the project report, in effect, giving the reader an overview of what is to come in the project report.

**CHAPTER 2: SURVEY OF TECHNOLOGIES**

**CHOICE OF PROGRAMMING LANGUAGE:**

While searching for the best programming language for our project, we encountered multiple languages that can fulfill our demand for analytical and scientific computing. These languages were having there own pros and cons. These languages were Python, R, Matlab, Julia, Java, Scala or C++. After carefully weighing them, we chose Python and here are our views in favor of the choice:

R and Matlab both are good scientific languages and they solve our purpose, but they don’t integrate easily into the production environment as Python. R is mainly used for statistical analysis and it is overwhelming for beginners, on the other, hand Matlab is very powerful almost for everything but it is costly, closed-source and needs a bit extra computer resource which is not feasible for us.

Julia is a young and powerful programming language that was developed while aiming specifically at “scientific computing, machine learning, data mining, large-scale linear algebra, distributed and parallel computing” [7]. But it has it’s own demerits like, arrays are 1-indexed, the language itself is immature. On the other, hand Python has large third-party package support which is indeed helpful.

Scala, Java or C++ can’t offer the same scientific libraries and community support as Python. Moreover, code written in Python is simple and offer more power while having very simple syntax.

So, Python being a general-purpose and a high-level language with vast community support and a solid ecosystem of scientific libraries is the best choice for us.

Here is a graph from the popular KDnuggets which efficiently plots the popularity of programming languages among the employers in the field of Machine Learning and Data Science:

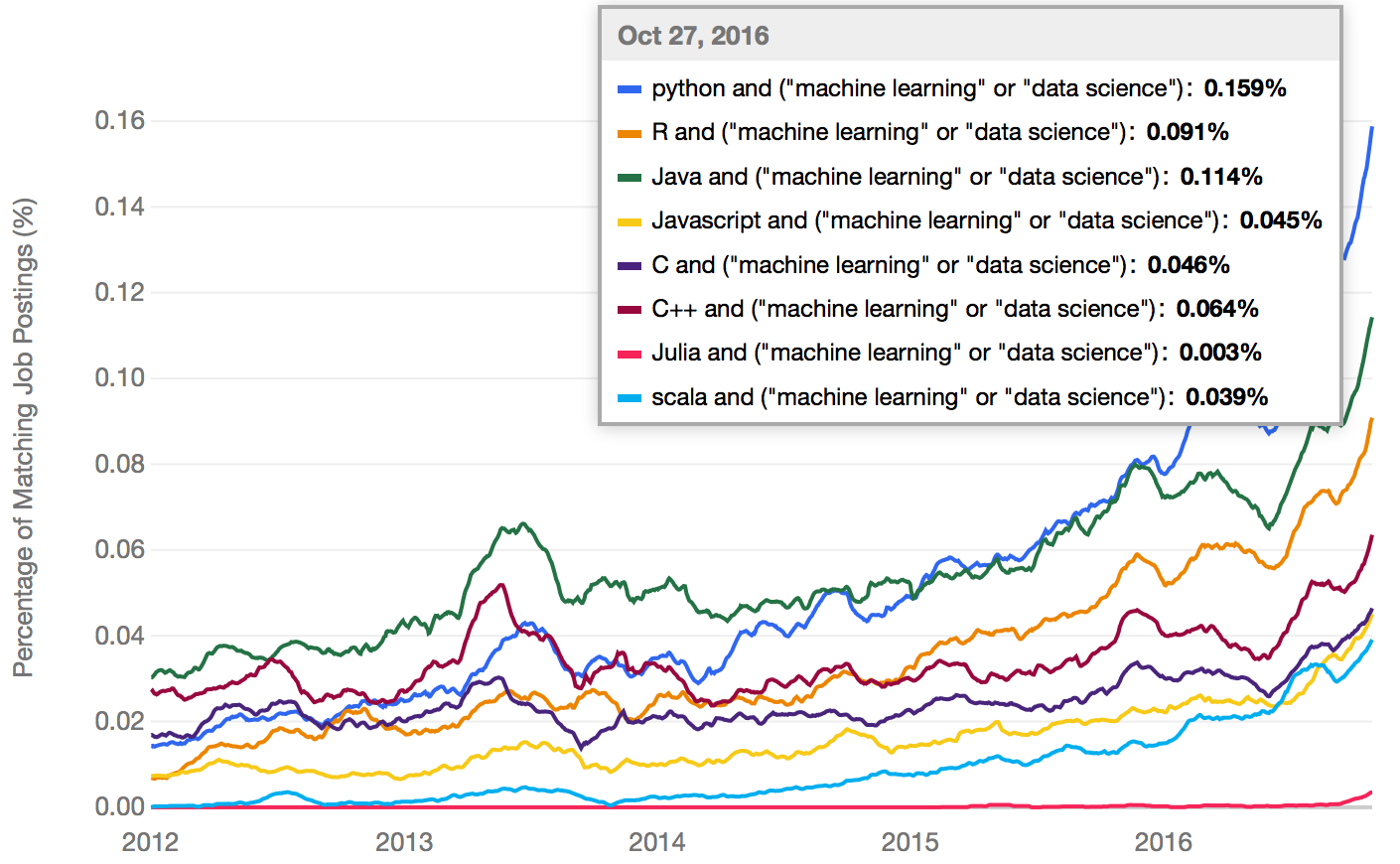


Figure 2. Skills employers are seeking [8]

Figure 3 is a graph representing the performance of NumPy over Pandas:

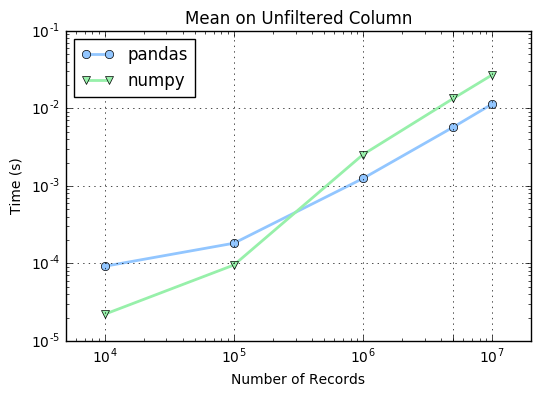


Figure 3: NumPy vs Pandas Performance [9]

Figure 4 is a relationship diagram representing the relation between Python and the libraries we are using:

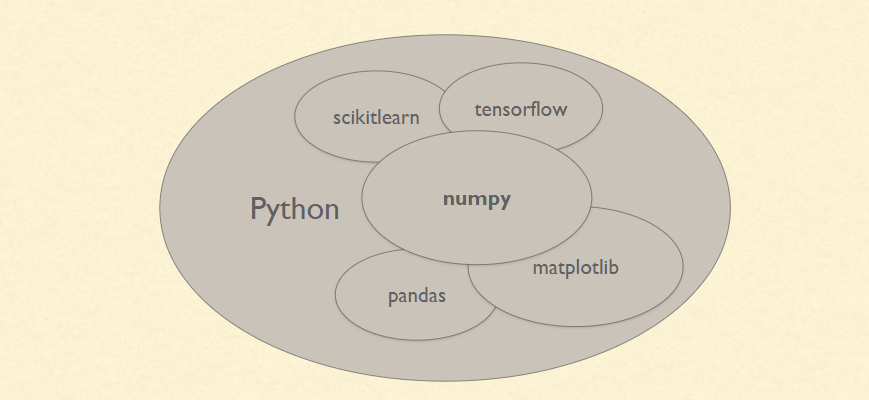


Figure 4: Python and various libraries [10]

**CHAPTER 3: REQUIREMENTS AND ANALYSIS**

**Problem Definition:**

In our project we try to determine the genre of a book from its summary by using the concepts of multi-label text classification. Books are not one dimensional because a book can span several genres. Thus we first set our multi-label classification problem. Book summaries are going to use different forms of a single word, mainly because to abide grammar of that particular language in which the document is written. In order to achieve higher precision, it becomes necessary to reduce such words to a common base form. This process is termed as stemming. So we implement stemming at the initial stage as it would allow the classifier to give a better accuracy , as it would give the classifier to give the ability to treat different forms of the same word as the same concept. Before building a classifier to perform the task of classification we removed the stopwords . Stopwords are the most common words which are used in any document and they carry far less meaning than other keywords in the text.After that we use a strategy to build a book genre prediction based on the summary and finally we create an inference function so that when new book summaries come in the future, our prediction system will be able to take the book summary as input and generate appropriate genre labels.

**Requirements Specification:**

1. Python programming environment is the first and foremost need of the system. Python is an object-oriented high-level programming language that is interpreted. It uses dynamic semantics. It incorporates dynamic data structures and other dynamic components which accelerate the overall programming development. The compilation free environment of Python makes it, even more, friendlier and faster for developers. Being an open-source and free language it is very much popular among the developers and even with beginners. It has its own huge community and hence vast support. Python is a general-purpose language and hence it can be used to achieve numerous objectives. It can be used in web technology, network technology, machine learning or artificial intelligence, data mining or even as a glue language to connect two different components and many more. Python has huge library support and many 3rd party packages which may count up to 20k! We are using Python for these numerous capabilities and simplicity to develop a powerful project.
2. Python dependent libraries like NumPy (Numeric Python) are a very powerful mathematical package that can perform high-level mathematical operations. Our program will use it to get access to the higher multi-dimensional array processing capability. Pandas is another package which is used for the powerful 2D data structure support. Though the processing speed and performance of Pandas is significantly less than NumPy we will be using it for the data analysis tools it offers. The data analysis tools Pandas offer is compact and easier to understand. Matplotlib is a 2D plotting library that helps us to get a visual representation of the data and we all know that the visual representation is much easier and has a higher appeal to it over the simple numeric representations or tabular representation. So, we will be using it to get intermediate results and also the final result in bar graphs or other plotting techniques. NLTK is an all-in-one natural language processing toolkit. Natural language processing is a very important step in text preprocessing. We will be utilizing the package NLTK for cleaning the text by removing whitespaces, backslash-apostrophe, everything except alphabets and converting everything to lowercase. It will also give us different frequencies like frequency of genre, frequency of words, etc. After this, it will help us to apply one of the stemming algorithm known as Porter stemmer and then remove stopwords. Scikit-Learn provides powerful yet efficient python modules for machine learning and data mining. We will be using it to convert the text into tf-idf features and build models. Scikit-Learn will also help us in building the inference function to update the training model in case a new and unique result comes up.
3. An integral part of our project requirement is the dataset. Every project related to text analysis and machine learning uses one or more datasets to train itself. Hence it is very important to acquire a good dataset. Generally, we prefer datasets with a huge number of data to achieve higher accuracy while training our model. Datasets may contain unnecessary or redundant values, so datasets require thorough cleanup via multiple approaches to reduce unnecessary resources wastage, computation power or complexities. We are using CMU Book Summary Dataset. This dataset contains plot summaries for 16559 books which were extracted from Wikipedia, along with aligned metadata from Freebase, including book author, title and genre [11]. This dataset contains unnecessary or redundant values, so we will be doing thorough cleanup via multiple approaches to reduce unnecessary resources wastage, computation power or complexities, in turn, this will generate a model with higher accuracy.
4. Finally, we will require an efficient integrated development environment (IDE), Spyder is our IDE which comes preinstalled within Anaconda. Also, everything related to the work is already in the system and we will be importing it whenever required. Spyder provides a text editor, an explorer or help area and a terminal. It provides good performance and access to all of the Anaconda Python library. So, this completes our working system.

The existing system upon which we are modifying lacked stemming operation, that’s why being a multiclass classification the system often produced an erroneous result and lacked accuracy. We are performing multilabel classification with the addition of stemming to produce result with a higher accuracy.

**Planning and Scheduling:**

Project planning is one of the most important steps in developing a project. Project planning is a process of setting up the scopes of the project and elucidating the objectives and the steps required to obtain them. The end result of project planning is the Project Management Plan.

In a broader aspect we can say, project planning is the approach the team is taking to accomplish the project and the steps involved to it.

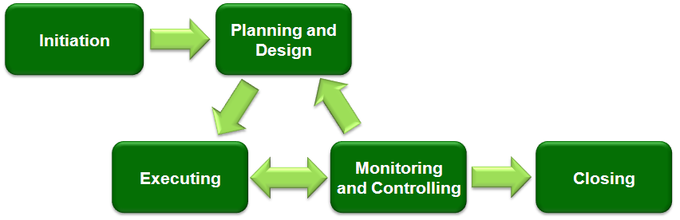


Figure 5: Steps in project planning [12]

So there are five major steps involved in the planning of a project:

**Project Initiation**: It is the starting phase of the project. This step tells us about the scope, objectives, output and the success criteria of the project. Within the initiation phase, the business problem or opportunity is identified, a solution is defined, a project is formed, and a project team is appointed to build and deliver the solution to the customer. A case consists of different features:

* A detailed description of the problem or opportunity with headings such as Introduction, Objectives, Problem Statement, Assumptions, and Constraints.
* A list of the alternative solutions available.
* A description of the preferred solution.
* Main project requirements.

So first a business case is designed. The business case should be easy to understand, clear, logical and relevant. The key aspects need to be tracked and measured and justified.

In our project “Text Classification using AI Techniques”, our project initiation phase consists of the tasks:

Firstly, the objective of our project is to classify a text or more precisely we can say a dataset to different classes using multiclass text classification approach. This project finds a bigger scope in the field of text classification or text categorization. We always need to classify a large set of data. As the library system is online and any other system where we have to keep a large set of data, classifying them is really needed. As we want to search an item, then depending upon the classification it becomes much easier.

Now talking about the output we can say that, after providing the large dataset to our model we get the output as classified depending upon our given parameters.

* **Project Planning and Design**: Project design is a phase of the project where a project's key features, structure, criteria for success are all planned out. Project design plays a very important role in the overall success of the project. Now, while planning a project a very important part is to keep the S.M.A.R.T goals in mind. These goals are:

**S**pecific – To set specific goals, answer the following questions: who, what, where, when etc.

**M**easurable – Creating the criteria that we can use to measure the success of a task.  
**A**ttainable – Identify the most important goals and what it will take to achieve them.  
**R**ealistic – You should be willing and able to work toward a particular goal.  
**T**imely – Create a time to achieve the goal.

The project planning step helps creating a proper plan which helps us execute the steps needed and brings us to the closure of the project. A project plan consists of many small plans like, creating a resource plan, a quality plan, a financial plan, a procurement plan etc. During this phase, the scope of the project is defined and a project management plan is developed. It involves identifying the cost, quality, available resources, and a realistic timetable.

* **Project execution**: The main focus of the project execution phase is to work on the actions or tasks designed on the design phase. The project execution phase tries to give the deliverables which are designed earlier. The project execution phase is the longest of all the phases and uses the maximum resources of all. During the project execution the execution team utilizes all the schedules, procedures and templates that were prepared and anticipated during prior phases. Unanticipated events and situations will inevitably be encountered, and the Project Manager and Project Team will have to deal with them as they come up. There are several tasks which are completed during the execution phase, some of them are:
* Assign resources.
* Execute project management plans properly.
* Update project modules.
* Modify plans as per needed.

In our project, the execution consists of the steps like: Collecting the dataset 🡪Cleaning it based on a different parameters 🡪 Train our model using the 80% of this dataset 🡪 Testing with the rest 20% dataset 🡪 Generating results.

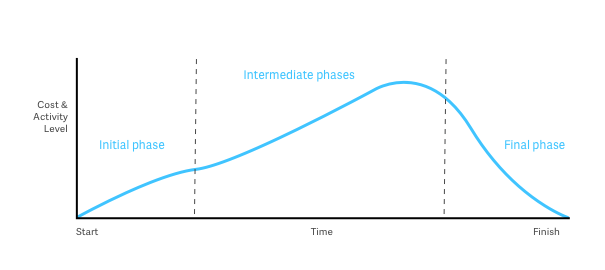
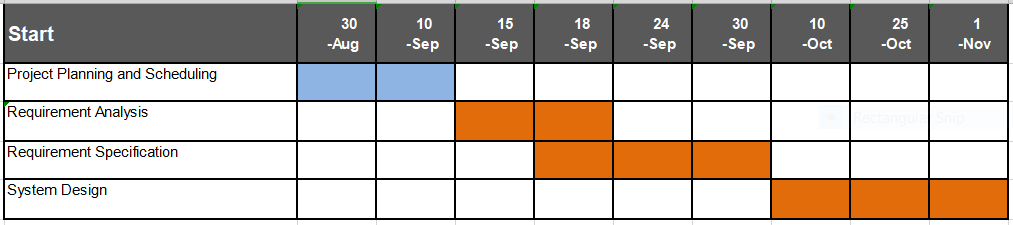
* **Monitoring and controlling:** The monitoring and controlling sees all the steps and metrices necessary or the project. Then it sees if the project can be completed within the given resources in terms of time, value and human resources. This process involves monitoring the observed output and the planned output. If we find a huge difference between this two, then some other approaches are used or otherwise we go on with same approach. Here in our project we try to find the genre of any book depending upon the summaries of those books. So if a book of genre ‘fiction’ is classified as something else then we need to monitor the output.
* **Project closing:** Then we come to the end of our project with the phase of project closing. In the project closing phase, all the final work is done. The finalization is got here in terms of everything. We should get the desired output and then the organizational team can be released to work on some other endeavors. Our project comes to an end with getting our desired output i.e. the book genre.

Figure 6: Relation between time and the major phases of a project [13]

Now, we are going to discuss about the constraints of our project. Those are:

* The dataset we are using should not contain null values. If there are null values, then, while training the data it may create problems.
* While extracting the main features of a big dataset, we should look into the most appropriate to avoid further ambiguities.
* As we know, big documents contain words which may have different meanings in different places. If we train the machine without giving proper description about them or removing them then it creates ambiguities.
* A large dataset may contain many columns and all of those may not come in our use. So while filtering the data in the first phase, these columns which we do not need, we should be removing all those.
* There should always be ways to improve our dataset for our use. Each time the dataset pass through a filter, the new dataset should be encapsulated with the training algorithm. Thus the accuracy of our project improves.



# Figure 7: Gantt Chart

**Software and Hardware Requirements:**

* + Hardware Requirement: Minimum of 384 CUDA cores GPU with 2GB memory, CPU with 3Ghz or higher, mouse, keyboard, 6GB of disk space, 8GB of RAM and a monitor.
  + Software Requirements: Windows 7 (or higher) or Linux v5 (or higher), Python, Anaconda, any text editor, a dataset of book-summary with tab-separated fields having at least one ID, book name, genre, summary, libraries (NumPy, Matplotlib, Pandas, Scikit-Learn, NLTK, JSON etc).

**Preliminary Product description:**

Our project consists of different steps. We have divided our whole task into several different subtasks. All these small subtasks yield different results and we get the final desired output.Now talking about the preliminary requirements of the system, we started searching for the best programming language and tool to use. Then we got MATLAB and R, both of them are very good and solve most of our purposes but do not fit properly into our production environment. Then Python comes into play and solves all of our purposes quite easily and accurately. Then after searching for a good IDE we got Anaconda, which is an IDE for python which comes with multiple editors, like-jupyter, spyder etc. From them, we are going to write our code in spyder editor which is a great editor. In the field of data science, python has found a lot of applications because of it’s enriched library. From the huge library of python , we are importing a few, like:

* NumPy: Due to its multidimensional array support for numerical computation and high-level mathematical functions.
* Pandas: Due to its powerful 2D data structure support and data analysis toolkit.
* NLTK: It is an all-in-one natural language toolkit.

Apart from these, we will use a multi-labelled dataset. The dataset, first of all should be labelled for use. Now as we know, a book can come under several genre, like- “adventure”, ”comedy”, ”drama”, so the book fits into all of these types. The dataset contains several columns, but after cleaning the dataset for getting lesser columns for our purposes.We also need the concept of Natural Language Processing here and with that we need tf-idf( term-frequency, inverse document frequency) tool. The primary objective of our project is to classify the genre of a book using multi-label text classification from its summary. We know one book can span multiple genres. So we have extracted the book summary and classified it into multiple genres.

The functionality of our project works like this: firstly, we have obtained a dataset and then we have started working on that dataset. In the dataset, firstly, we have separated the columns which are only needed for our project from the large dataset. Then, we have used the stemming approach which is Potter’s Stemmer for this project. Then have cleaned the dataset based on different parameters, like null values and occurrence of same words etc. and the model is trained using that dataset.

The final output of our project will be the classified genres which will be obtained after giving input the book summary dataset.

* **Conceptual Models:** The basic conceptual model of our project consists of several sub-tasks and those are:

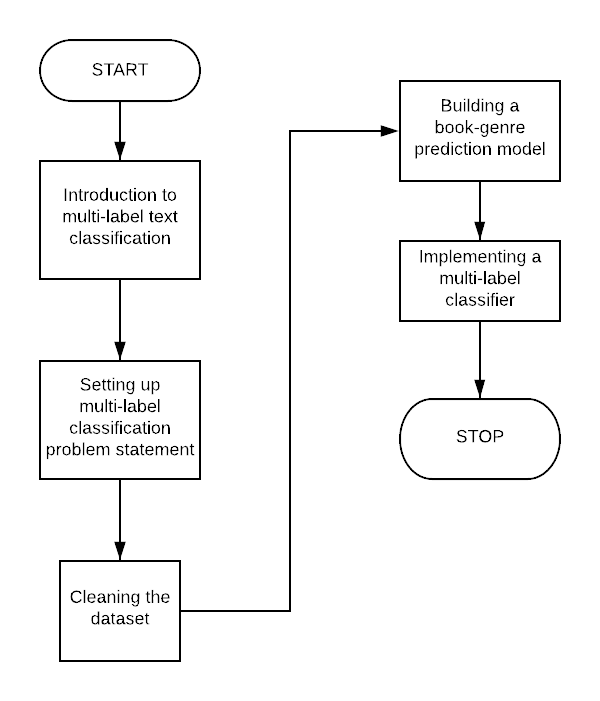
****

Figure 8: Conceptual model of our project

##### **CHAPTER 4: SYSTEM DESIGN**

##### **Basic Modules:** We are trying to predict book genre based on its summary using Artificial Intelligence (A.I) techniques. The task can be broken down into various manageable parts -

##### **Brief Introduction to multi-label classification:** The apparent concept behind multi-label classification is in the name itself, here an instance can have multiple labels and the number is not fixed unlike binary classifier where the number of classes is two, or multi-class classifier where there are more than two labels but only one label is there for each input.

##### **Setting up our Multi-label Classification problem statement:** Books are not one-dimensional , they span several genres. We extracted a bunch of book summaries and got down to work using concept of multi-label classification. After that we determine the number of book genres covered in the dataset.

##### **Cleaning the Dataset to make it ready for book genre prediction:** Book summaries are going to use different forms of a single word, mainly because to abide grammar of that particular language in which the document is written. In order to achieve higher precision, it becomes necessary to reduce such words to a common base form. This process is termed as stemming. We implement stemming at the initial stage as it would allow the classifier to give a better accuracy as it would allow the classifier to treat different forms of the same word as the same concept. After that we remove stop words from the summaries.

##### **Our strategy to build a book genre prediction model:** We split our data into train and validation sets for training and evaluating our model’s performance. We did it with a 80-20 split – 80% of the data samples in the train set and the rest in the evaluation then turned our focus of extracting features from the cleaned version of book summaries. For our project we have used tf-idf.

##### **Implementing a multi-label classifier to build a book genre prediction model:** We then build a regression model, and we choose a logistic regression model as it is quick to train. We then also take care of the new data or book summary that will come in the future. Our book genre prediction system should be able to take a book summary in raw from input and generate its genre tags.

##### **Data Design:**

**Schema Design:** We have used the CMU Book summary dataset. The dataset contains plot summaries for 16,559 books extracted from Wikipedia, along with aligned metadeta from Freebase, including book author, title and genre. The following example illustrates the data and metadata available for Don DeLillo’s White Noise.

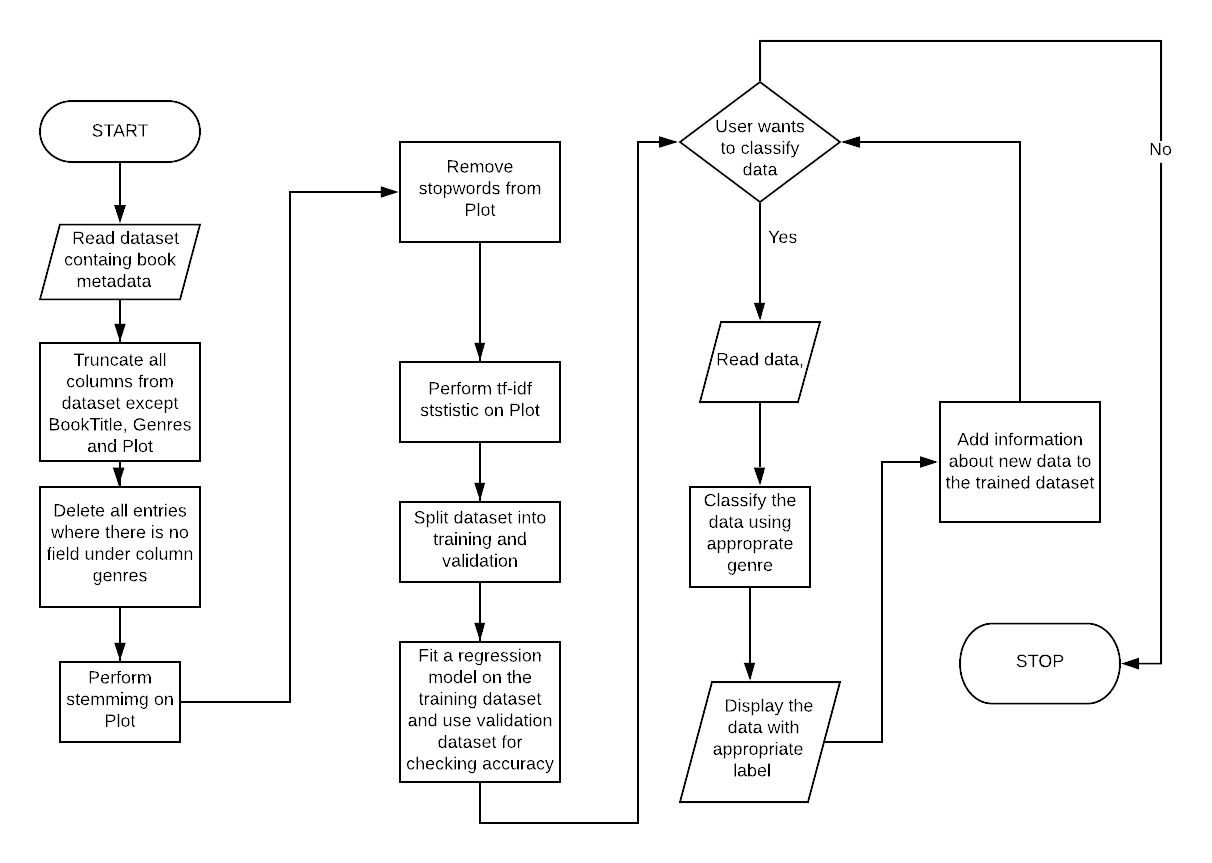
**Table 4 : Metadata available for Don DeLillo’s White Noise.**

|  |  |
| --- | --- |
| **Wikipedia ID** | 1166383 |
| **Freebase ID** | /m/04cvx9 |
| **Book title** | White Noise |
| **Book author** | Don DeLillo |
| **Publication date** | 1985-01-21 |
| **Genres** | Novel, Postmodernism, Speculative, Fiction |
| **Summary** | Set at a bucolic …...surviving. |

**Data Integrity and Constraints:** We only require the Book title, Genres and Plot from the dataset , rest of the columns are discarded. We removed all the unnecessary blank spaces, special characters from the dataset . Rows were Genres column is empty are discarded too.

**Procedural Design:** Procedural design is a systematic way for developing algorithms or procedurals.

* **Flowchart :**

****

**Figure 9:** Flow chart for the algorithm

**Algorithms Design:**  The proposed method consists of four connected sections, viz., Initialize, Cleaning, Training and Classify. Corresponding to each section, the algorithm is stated below:

**Algorithm: Initialize**

BEGIN

(a)Read the dataset containing book metadata.

(b) Truncate all columns from the dataset except Book Title, Genres and Plot.

(c) Delete all rows where there is no entry for the field under column Genres.

END

**Algorithm: Cleaning**

BEGIN

(a) Perform Stemming on the dataset in order to reduce the words used in Plot to their morphological root or stem.

(b)Remove the stop words from Plot.

(c)Perform tf-idf statistic on the Plot.

END

**Algorithm: Training**

BEGIN

(a)Split the dataset into two for training and validation respectively.

(b)Fit a regression model using the training dataset and use the validation dataset to check for accuracy.

END

**Algorithm: Classify**

BEGIN

LOOP while user wants to classify data

(a)Read data entered by user

(b)Preprocess the data as in the Initialize process

(c)Classify the new instance of data by assigning appropriate genres.

(d)Add information about the newly classified data instance and its label to the trained dataset

END LOOP

END

**CONCLUSION**

Text classification is the need of the hour. We are taking some small steps towards the improvement of this technique. Earlier this technique was used in previous works where the accuracy of prediction was not very high. But now, after using some great statistical tools and algorithms like logical regression we are getting some more accurate results. We are developing our model based on the logical-regression method. Some of the previous works on text classification were based on binary classification and multi-class classification. But working on the book summary dataset, we can see a book can span multiple genres, so we will be trying to implement multi-label classification in this project.For the time being we have tried to develop an algorithm which will be using our dataset and train machine learning models using it.

**REFERENCES**

1. Power, R., Chen, J., Karthik, T., & Subramanian, L. (2010). Document classification for focused topics. In Artificial Intelligence for Development - Papers from the AAAI Spring Symposium, Technical Report (Vol. SS-10-01, pp. 67-72)
2. Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development,1(4), 309-317.
3. A. Sarkar, S. Chatterjee, W. Das, D. Datta, “Text Classification using Support Vector Machine”, International Journal of Engineering Science Invention, Vol. 4 Issue 11, November 2015, pp. 33 – 37.
4. Anurag Sarkar, Debabrata Datta,"A Frequency Based Approach to Multi-Class Text Classification",International Journal of Information Technology and Computer Science(IJITCS), Vol.9, No.5, pp.15-22, 2017. DOI:10.5815/ijitcs.2017.05.03
5. S. Robertson, "Understanding Inverse Document Frequency: On theoretical arguments for IDF", Journal of Documentation 60 no. 5, pp. 503 – 520.
6. M. Ikonomakis, S. Kotsiantis, V. Tampakas, “Text Classification Using Machine Learning Techniques”,WSEAS Transactions on Computers, Vol. 4 Issue 8,August 2005, pp. 966 – 974.
7. https://www.infoworld.com/article/3241107/julia-vs-python-which-is-best-for-data-science.html (Last accessed: 12/10/2019 13:04)
8. https://www.kdnuggets.com/2017/01/most-popular-language-machine-learning-data-science.html (Last accessed: 14/10/2019 18:23)
9. https://docs.anaconda.com/anaconda/ (Last accessed: 19/10/2019 9:56)
10. http://gouthamanbalaraman.com/blog/numpy-vs-pandas-comparison.html (Last accessed: 19/10/2019 23:17)
11. https://cloudxlab.com/blog/numpy-pandas-introduction/ (Last accessed: 26/10/2019 19:54)
12. www.cs.cmu.edu/~dbamman/booksummaries.html (Last accessed: 31/10/2019 07:09)

[13] [https://canvas.uw.edu/courses/1113257/pages/project-management-phases](https://canvas.uw.edu/courses/1113257/pages/project-management-phases" \t "https://mail.google.com/mail/u/0/_blank) (Last accessed: 01/11/2019 22:00)

[14][https://courses.lumenlearning.com/boundless-management/chapter/the-planning-cycle/](https://courses.lumenlearning.com/boundless-management/chapter/the-planning-cycle/" \t "https://mail.google.com/mail/u/0/_blank) (Last accessed: 01/11/2019 23:27)