DYNAMIC NEWS CLASSIFICATION

Enrollment Number	-	9913103503
Name of Student	-	Ashutosh Khurana
Name of Supervisor	-	Mr. Sudhanshu Kulshrestha

`

May – 2017

Submitted in partial fulfillment of the Degree of

B.Tech or
5 Year Dual Degree Programme B. Tech

in

Computer Science Engineering

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING & INFORMATION TECHNOLOGY

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

TABLE OF CONTENTS

DECLARATION

We thusly proclaim that this thesis is my/our own particular work and that, to the best of my insight and conviction, it contains no material beforehand distributed or composed by someone else nor material which has been acknowledged for the honor of some other degree or certificate of the college or other establishment of higher learning, with the exception of where due affirmation has been made in the content.

Place: JIIT, Noida Signature:

Date: 13/5/2017 Name : Ashutosh Khurana

Enrollment No : 9913103503

CERTIFICATE

This is to guarantee that the work titled "Dynamic News Classification" put together by "Ashutosh Khurana" is in complete satisfaction for the honor of level of B.Tech of JIIT, Noida has been completed under my supervision. This work has not been submitted somewhat or entirely to whatever other University or Institute for the honor of this or some other degree or recognition.

Signature of Supervisor
Name of Supervisor Mr. Sudhanshu Kulshrestha
Date 13/5/2017

ACKNOWLEDGEMENT

I want to put on record our profound feeling of appreciation for Professor Sudhanshu Kulshrestha for liberal direction, help and valuable proposals. I want to express gratitude toward Jaypee Institute of Information Technology, Noida, for its priceless direction and help, without which the achievement of the errand would have never been conceivable. I likewise express gratitude toward them for giving this chance to investigate into this present reality and understand the interrelation of theoretical concepts and their applications. I likewise wish to extend my gratitude to various colleagues for their quick remarks and useful recommendations to enhance the nature of this detailed work.

Signature of the Student	
Name of the Student	Ashutosh Khurana
Enrollment Number	9913103503
Date	13/5/2017

LIST OF FIGURES

LIST OF TABLES

1. Introduction

1.1. Introduction

In today's world, the ability to automatically classify documents into a fixed set of categories is highly desirable. Common scenarios include classifying a large amount of unclassified archival documents such as newspaper articles, legal records and academic papers. The exponential growth of the data may lead us to a time in future where huge amount of data would not be able to be managed easily. Text Classification is done through Text Mining study which would help sorting the important texts from the content or a document to manage the data or information easily.

Example: When there are large amount of unstructured data containing information about Politics, Sports, Health, Technology etc. and all of these are scattered and the user wants to look at a specific category like Sports so it would become very difficult for him to look through all of the data.

There comes use of the Classifier, to assign a topic to each and every article so that there becomes a uniformity of the above topics and user's job to find an article becomes less cumbersome and less monotonous. Natural language processing offers powerful techniques for automatically classifying documents. Salient features for document classification may include word structure, word frequency, and natural language structure in each document.

The report includes an in-depth analysis of automatic classification of news and daily information articles into six pre-defined buckets - business, politics, technology, sports, entertainment, and health. We have collected about 3600 articles with 600 in each of the following 6 categories. We have used web scraping technique to collect the dataset. And used Naive bayes, Latent Dirichlet Allocation, Max entropy,K-NN & Decision Tree techniques to classify the news articles into predefined categories.

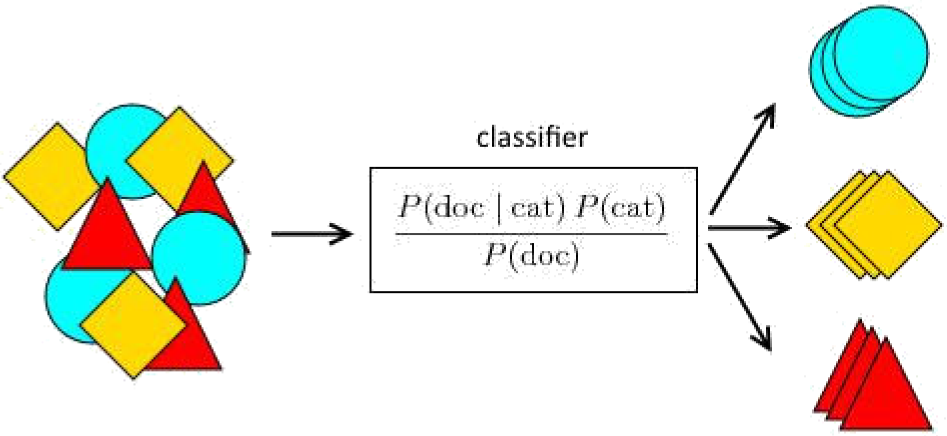

classifier
$\frac{P(\mathrm{doc} \mid \mathrm{cat})\, P(\mathrm{cat})}{P(\mathrm{doc})}$

Figure 1-Classification

1.2. List some relevant current/open problems

Data classification and identification is all about tagging your data so that it can be found quickly and efficiently. Data classification helps in deleting duplicate data which in turns helps cut storage and backup costs for an organization. Classification can help an organization to meet legitimate and administrative prerequisites for recovering particular data inside a set time span, and this is frequently the inspiration driving executing data classification technology. In any case, information procedures contrast significantly starting with one association then onto the next, as each creates diverse sorts and volumes of information. The adjust may change incredibly starting with one client then onto the next between office reports, email correspondence, pictures, video records, client and item data, money related information and so on. It might appear a smart thought to group and tag everything in the databases.

1.3. Problem Statement

Due to exponential growth of data it has become a tedious job to search all over the data to find a particular article. Hence we use data classification so that we find the required article quickly and efficiently. Data classification helps cut storage cost by reducing redundant data. For an organization to work efficiently and properly it is important that they keep their data in classified manner. Major problem we are facing now days is the classification of news articles as news data is growing at an exponential rate.

We have used two learning techniques – supervised and unsupervised – to classify the data in the articles. Supervised learning techniques include Naïve Bayes in which we use train data and testing data to create a model. Unstructured data gets classified using unsupervised learning technique. Latent Dirichlet Allocation (LDA) is one such example of unsupervised technique which takes random shuffled data as input and classifies the data into relevant topics.

1.4. Overview of Proposed Solution approach and Novelty/benefits

We have used four machine learning techniques to classify news articles dynamically. These include naïve bayes, TF-IDF, max entropy and Latent Dirichlet Allocation. LDA comes under unsupervised while remaining of these comes under supervised learning technique. Each of these techniques can be used to classify the data but there are different scenarios or conditions under which each of these proposed techniques give optimum results. The naive Bayes is simpler and easy to understand than other supervised learning. We can use naive bayes if we have less training data. Naives Bayes technique will converge at a faster rate if conditional independence holds true.

The most relevant terms, as detected by complex algorithms, are extracted by TF-IDF. The similarity between different documents can be easily computed by Tf-Idf.We use LDA when we have a large amount of unstructured data. And we simply want to classify the unstructured data into some relevant topics according to document topic and topic word probabilities.The Maximum Entropy requires very carefully selected feature sets else performance can be very poor. Like naive bayes, we can use maximum entropy if we have less training data.

1.5. Literature Survey

- Paper 1:

  Title- **Document Classification for Newspaper Articles**

  Citation- Ramdass, Dennis, and Shreyes Seshasai. "Document classificationfor newspaper articles." (2009).

- Paper 2:

  Title- **Online News Classification**

  Citation- Kaur, Harmandeep, Sheenam Malhotra, and Fatehgarh Sahib."Online News Classification: A Review." *International journal of Innovationin Engineering and Technology (IJIET)* 2.2 (2013).

- Paper 3:

  Title: **News Classifications with Labeled LDA**

  Citation- Bai, Yiqi, and Jie Wang. "News classifications with labeled LDA."

  Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on. Vol. 1. SCITEPRESS,2015.

- Paper 4:

  Title: **An Intelligent Financial Web News Articles Digest System**

  Citation- Lam, Wai, and Kei Shiu Ho. "FIDS: an intelligent financial Webnews articles digest system." *IEEE Transactions on Systems, Man, andCybernetics-Part A: Systems and Humans* 31.6 (2001): 753-762.

- Paper 5:

  Title: **Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency.**

Citation- Hakim, Ari Aulia, et al. "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach." *Information Technology and ElectricalEngineering (ICITEE), 2014 6th International Conference on*. IEEE, 2014.

- Paper 6:

  Title: **Learning Approaches for Detecting and Tracking news Events**

  Citation- Yang, Yiming, et al. "Learning approaches for detecting andtracking news events." (2000).

- Paper 7:

  Title : **Text Categorization of KNN Based on K-means for Class Imbalanced Problem**

  Citation- W. Yu and X. Linying, "Research on Text Categorization of KNN Based on K-Means for Class Imbalanced Problem," *2016 Sixth InternationalConference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC).*

- Paper 8:

  Title **Urdu Text Classification**

  Citation : 12th International Conference on High-capacity Optical Networks and Enabling Emerging Technologies (HONET)*, Islamabad, 2015, pp. 1-4.*

- Paper 9:

  Title : **Machine Learning techniques for Data Mining : A Survey**

  Citation: IEEE International Conference on Computational Intelligence and Computing Research, Enathi, 2013, pp. 1-6.

1.5.1. Summary of papers

1.

Title Of Paper	**Classification for Newspaper Articles**
Authors	Ramdass, Dennis and Shreyes Seshasai
Year Of Publication	2009
Publishing Details	6.863 Final Project Spring
Summary	In this research paper [1], they talk about the capacity to naturally group data into a settled arrangement of classes. For instance, daily paper articles can be named 'highlights', "games" or 'news'.They utilize the NLP (Natural dialect preparing) strategies for consequently grouping documents.They try to group the daily paper articles from the MIT daily paper The Tech. ALL daily paper articles from the MIT daily paper The Tech divide into six diverse class labels, for example, Arts, Features, News, Sentiment, World and Sports. Administered learning occurred on a chronicle of 3000 articles with 500 articles from each of the 6 classes. These articles were the most as of late distributed 500 articles in every class. They haphazardly split this chronicle of ordered records into preparing and testing bunches for the characterization frameworks alluded as classifiers. The Research Paper mostly center in the zone of Naive Bayes characterization, Maximum Entropy classification and probabilistic linguistic use characterization. They utilized the Natural Language Toolkit (NLTK) bundle to execute every one of the three classifier in Python.The research paper analyzes machine learning techniques and lists representation strategies. An examination of highlight choice strategies and arrangement calculations were introduced. Among the three the best classifier utilized multi-variate Bernoulli highlights with the guileless bayesian classifier with a precision of 76%.
Weblink	http://web.mit.edu/6.863/www/fall2012/projects/writeups/newspaper-article-classifier.pdf

2.

Title Of Paper	**Online News Classification**
Authors	Kaur, Harmandeep, Sheenam Malhotra, and Fatehgarh Sahib
Year Of Publication	2013
Publishing Details	International journal of Innovation in Engineering and Technology (IJIET) 2.2
	In this paper [2],they give the introduction of text arrangement, procedure of content characterization and additionally the review of the classifiers and look at some current classifier on the premise of couple of criteria like time complexity, principal and performance.A effective execution has been demonstrated to in this paper like proper methodologies to extricate the news from the online entryways and further processing the data with techniques like Tokenization, stemming word, removing stop words.They utilize two model HMM and SVM. The paper details out how Hidden Markov Model is used for content extraction and Support Vector Machine, the other model, can be utilized for characterization. They additionally utilize K mean calculation to make the group and CART calculation to speak to it into various leveled shape
Web Link	http://web.mit.edu/6.863/www/fall2012/projects/writeups/newspaper-article-classifier.pdf

3.

Title Of Paper	**News Classifications with Labeled LDA**
Authors	Kaur, Harmandeep, Sheenam Malhotra, and Fatehgarh Sahib
Year Of Publication	2015
Publishing Details	Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on. Vol. 1. SCITEPRESS
Summary	In this paper [3], they have shown a comparison between labeled LDA (Latent Dirichlet Allocation) and SVM(support vector machine) [10] using a chinese dataset.There are 2 types of learnings on the basis of which documents can be classified which are Supervised and Unsupervised Learning. LDA is an unsupervised machine learning technique.They have used 80% of the dataset as training data and remaining 20% of the dataset as testing data. Total of 5000 articles were used as dataset with 10 categories namely Politics, Technology ,Military, Sports, Entertainment, Health, History, Real estate, Automobiles and Games.Result of the comparison between the two machine learning techniques include that for all classifier larger training set will produce higher accuracy and LLDA has higher precision than SVM.
Web Link	http://ieeexplore.ieee.org/abstract/document/7526905/

4.

Title Of Paper	**An Intelligent Financial Web News Articles Digest**
Authors	Lam, Wai, and Kei Shiu Ho
Year Of Publication	2002
Publishing Details	IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 31.6 (2001): 753-762.
Summary	In this paper [4], they have outlined a framework called FIDS (Financial News Digest System) which can process on the web monetary news naturally. The system perform three tasks i.e classification, data extraction and data enquiry. News articles are intermittently downloaded from on the web sources. In light of its substance each article is ordered into one of the five classifications to be specific organization execution, economy, merger and obtaining, administrations and items, and securities. For every classification a template is pre characterized. Template comprises of rundown of data which one can expect (outline) from a standard article in that class. Data removed from the article is then filled in the layout to be spoken to as synopsis to the end client and can be put away in the database for additionally inquiry. The data enquiry subsystem bolsters client inquiries on the committed classification information put away in the money related data database. In this manner the uniqueness of FIDS lives in the way that it can address various areas at the same time.
Web Link	http://ieeexplore.ieee.org/abstract/document/7526905/

5.

Title Of Paper	**Automated Document Classification for News Article based on TF-IDF**
Authors	Hakim, Ari Aulia
Year Of Publication	2014
Publishing Details	Information Technology and Electrical Engineering (ICITEE), 2014 6th International Conference on. IEEE, 2014.
Summary	In this paper [5] , the writer has classified online news article utilizing Term Frequency–Inverse Document Frequency (TF-IDF) algorithm.12,000 articles were accumulated and 53 people were to physically aggregate the articles on its points. PC took 151 hours to actualize the entire methodology totally and it was done utilizing Java Programming Language. The precision of this classifier was 98.3 % .The inconveniences of utilizing this classifier was it required a considerable measure of investment because of expansive number of words in the word reference. Now and then the content contained a ton of words that depicted another classification since the calculation considers each word's weight and made the framework make a wrong yield. This classifier was utilized basically for 2 reasons which were it is one of the most perceived word weighting calculations and exactness of the above classifier is promising as it uses 2 approaches Term Recurrence and Inverse Document Frequency.
Web Link	http://ieeexplore.ieee.org/document/7007894/

6.

Title Of Paper	**Learning Approaches for Detecting and Tracking news Events**
Authors	Yang, Yiming
Year Of Publication	2000
Publishing Details	Journal IEEE Intelligent Systems Volume 14 Issue 4, July 2000
Summary	The goal of this research paper [6], is to Detect and Track News events (TDT). Authors have used manually segmented documents in this paper. Event Detection or Topic Detection is done using 2 methods which are Retrospective and Online Detection which is done using document clustering. Different methods are used for the document clustering and to calculate the weight of the words present in the former , used was TF-IDF(Term Frequency and Inverse Document Frequency).Event Tracking assign topics to novice news stories based on the previous identified past stories that define the event. k-NN and Decision Tree Induction were the two important algorithms used in this paper. These are used to classify the documents which are detected by Detection process mentioned above. The authors found out that Retrospective detection was better than Online Detection as in the latter,Non-Clustering approaches showed better accuracy than clustering. However according to them, this required further investigation. Document Clustering can be useful only if event identification takes user input into account and users should get suggested browsing strategies along with system generated clusters. K-NN and Decision tree showed good performance for positive training examples.
Web Link	http://repository.cmu.edu/cgi/viewcontent.cgi?article=1325&amp;context=compsci

7.

Title Of Paper	Text Categorization of KNN Based on K-means for Class Imbalanced Problem
Authors	Wang Yu and Xu Linying
Year Of Publication	2016
Publishing Details	Sixth International Conference on Instrumentation & Measurement, Computer, Communication and Control
Summary	The problem of Class Imbalancing has been discussed in detail in this research paper. Information are said to endure the Class Imbalance Problem when the class distributions are exceedingly imbalanced and this prompts lessening of the exactness of the classifier.The creators have utilized four datasets which have diverse class imbalanced rates are extricated from the whole corpus.KNN calculation, an enhanced version, in light of K-means clustering has been proposed as the most effective solution in this paper. The enhanced calculation has higher order exactness and solidness and it is reasonable for the dataset with a wide range of dispersion, particularly for the dataset with class imbalanced issue. The relative investigation of exploratory outcomes demonstrates that for an imbalanced dataset,contrasted with the exemplary K-NN calculations, NWKNN3 and NWKNN5, K means-KNN calculation has a higher order precision.
Web Link	http://ieeexplore.ieee.org/document/7774847/

8.

Title Of Paper	Urdu Text Classification
Authors	K.Khan,R. Ullah khan, Ali Alkhalifah, N. Ahmad
Year Of Publication	2015
Publishing Details	12th International Conference on High-capacity Optical Networks and Enabling Emerging Technologies (HONET), *Islamabad, 2015, pp. 1-4.*
Summary	In this paper, the text classification of Urdu language is done. The authors used pre-processing,feature extraction and Decision Tree Classifier.The aforementioned feature extraction techniques can be further sub-classified into Principal Component Analysis, Hu Moments and Zernike Moments.There were no database or corpus for Urdu characters,so as a first step , a database of the former was created.The database consisted of 441 characters. An accuracy of 92.06 % was secured by Hu moment.
Web Link	http://ieeexplore.ieee.org/document/7395445/

9.

Title Of Paper	Machine Learning techniques for Data Mining : A Survey
Authors	S. Sharma, J. Agrawal, S. Agarwal and S. Sharma
Year Of Publication	2013
Publishing Details	IEEE International Conference on Computational Intelligence and Computing Research, Enathi, 2013, pp. 1-6.
Summary	In this paper, creators have exhibited different classifier strategies including Decision Tree, Support Vector Machine, Nearest Neighbor. Various calculations are consolidated for arranging the informational index. This paper gives compressive review of different arrangement strategies utilized as a part of various fields of information mining.
Web Link	http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6724149&isnumber=6724108

1.5.2. Integrated summary of the literature studied

While going through all the 9 research papers we came across many classification techniques namely Naive Bayes, Decision Tree, K-Nearest Neighbor, Latent Dirichlet Allocation, Maximum Entropy and TF-IDF and all of these techniques have their own pros and cons in different situations. Naive bayes, Decision Tree and K-NN are Supervised learning techniques while Latent dirichlet Allocation(LDA) is unsupervised learning technique. TF-IDF is used to count the weight of the word in each document.

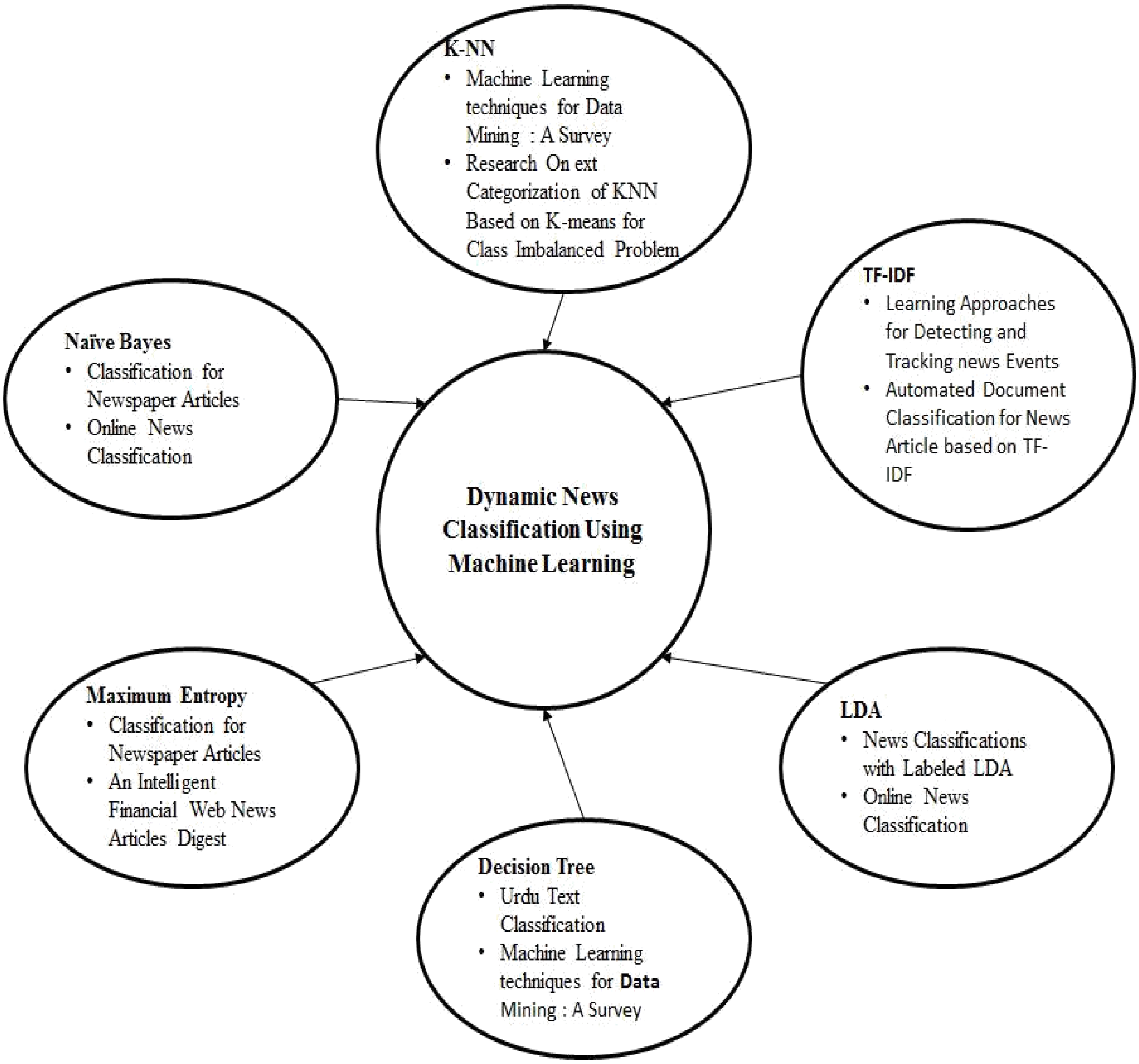

K-NN
• Machine Learning techniques for Data Mining : A Survey
• Research On ext Categorization of KNN Based on K-means for Class Imbalanced Problem
TF-IDF
• Learning Approaches for Detecting and Tracking news Events
• Automated Document Classification for News Article based on TF-IDF
Naïve Bayes
• Classification for Newspaper Articles
• Online News Classification
Dynamic News Classification Using Machine Learning
Maximum Entropy
• Classification for Newspaper Articles
• An Intelligent Financial Web News Articles Digest
LDA
• News Classifications with Labeled LDA
• Online News Classification
Decision Tree
• Urdu Text Classification
• Machine Learning techniques for Data Mining : A Survey

Figure 2 - Integrated Summary of Literature Studied

1.6. Details of Empirical Study

Natural Language Toolkit- NLTK is one of the leading platforms for designing programs based on Python to interpret and analyze human language data. Various datasets and resources like Word-net employ NLK for their interfaces and spectrum of text processing features like tokenization, parsing, classification etc.

Scikit-Learn –It is a package in the Python with set of advanced classifiers. It accommodates various classifications, regression and clustering algorithms, gradient boosting, K-means etc and is drafted to inter-operate with python numerical and scientific libraries NumPy & SciPy respectively. It is predominantly written in Python, with some of the algorithms written in Cython.

Beautiful Soup (HTML parser)- It is a python package employed to extract data , or rather , parse HTML and XML documents. It generates a parse tree for parsed pages which can then be utilized for extracting data from HTML, which is eventually used for web scrapping.

2. Analysis, Design and Modeling

2.1. Requirement Specification

I. For Windows

A. Operating System

- Windows 7 and above
- Windows Server 2008 and above

B. Processors

- Any Processor
- AVX2 instruction set support is needed
- 4 cores is recommended with Polyspace

C. Disk Space

- MATLAB - 2 GB
- Custom installation – 4-6 GB

D. RAM

- Default - 2 GB
- Simulink option - 4 GB
- Polyspace option - 4 GB for every core

E. Graphics

- Graphics card is not a mandatory requirement.
- Hardware accelerated graphics card

For MAC

A. Operating System

- macOS Yosemite and above

B. Processor

- Any Intel x86-64 processor
- AVX2 instruction set support is recommended
- With Polyspace, 4 cores is recommended

C. Disk Space

- 2 GB for MATLAB only
- 4 - 6 GB for a typical installation

D. RAM

- Default - 2 GB
- Simulink option - 4 GB
- Polyspace option - 4 GB for every core

E. GRAPHICS

- Graphic card is not a mandatory requirement
- Hardware accelerated graphics card

2.2. Functional and Non Functional requirements

2.2.1. Functional Requirement

- Classify news articles in predefined 6 categories.
- Scrap data from websites to obtain training and test data
- Removing stop words from dataset
- Natural Language Toolkit(NLTK)
- Sciket-learn
- Beautiful Soup

2.2.2. Non Functional Requirement

- Quantity of train/test dataset
- Classification categories
- Time required to execute the algorithm
- Quantity of input data to the model
- Processor on which model is running
- Minimum ram required for running the model

2.3. Overall architecture with component description and dependency details

We have designed a model in which we are classifying news articles based on the following categories:

- Business/Economy
- Entertainment
- Health
- Politics
- Sports
- Technology/Science

We have used both types of learning techniques - supervised (Naive Bayes and Maximum Entropy) and unsupervised (Latent Dirichlet Allocation) to classify the news articles. We have used dataset from bbc, the guardian and reuters.

Component description

Anaconda- It is an open source distribution of R plus Python programming languages, primarily used for extensive data processing and concepts the likes of predictive analysis and scientific computing. Its principle aim is to make package management and deployment easier and more comprehensible. The package management system it puts into use is called **Conda** and it is a freemium product, that is, basic services are provided free of charge whereas more advanced versions need to be paid for. Even though it is mostly written in Python, some of the modules have been codified in C. This software offers its users two modes; a text mode and a GUI, making it convenient to install on a wide range of systems.

- **Natural Language Toolkit**- Developed by Steven Bird and Edward Loper, NLTK is essentially a suite of libraries and programs put to use for statistical and symbolic Natural Language Processing of written English in language of Python. This toolkit aspires to assist research and teaching in NLP and similarly related concepts of empirical linguistics, artificial intelligence, cognitive science and machine learning.NLTK supports classification, tokenization, stemming, tagging, parsing, and semantic reasoning functionalities and includes graphical demonstrations and sample data.

- **Scikit-learn-** Scikit was developed to be used in Python. It accommodates various classifications, regression and clustering algorithms, gradient boosting, K-means etc and is drafted to interoperate with the python numerical and scientific libraries such as NumPy and SciPy. In addition to the major portion written in Python, a small subset of algorithms is written in Cython.

- **Beautiful Soup (HTML parser)** - It is a Python package employed to extract data, or rather, parse HTML and XML documents. It generates a parse tree for parsed pages which can then be utilized for extracting data from HTML, which is eventually used for web scraping.

2.4. Design Documentation

2.4.1. Use Case diagrams

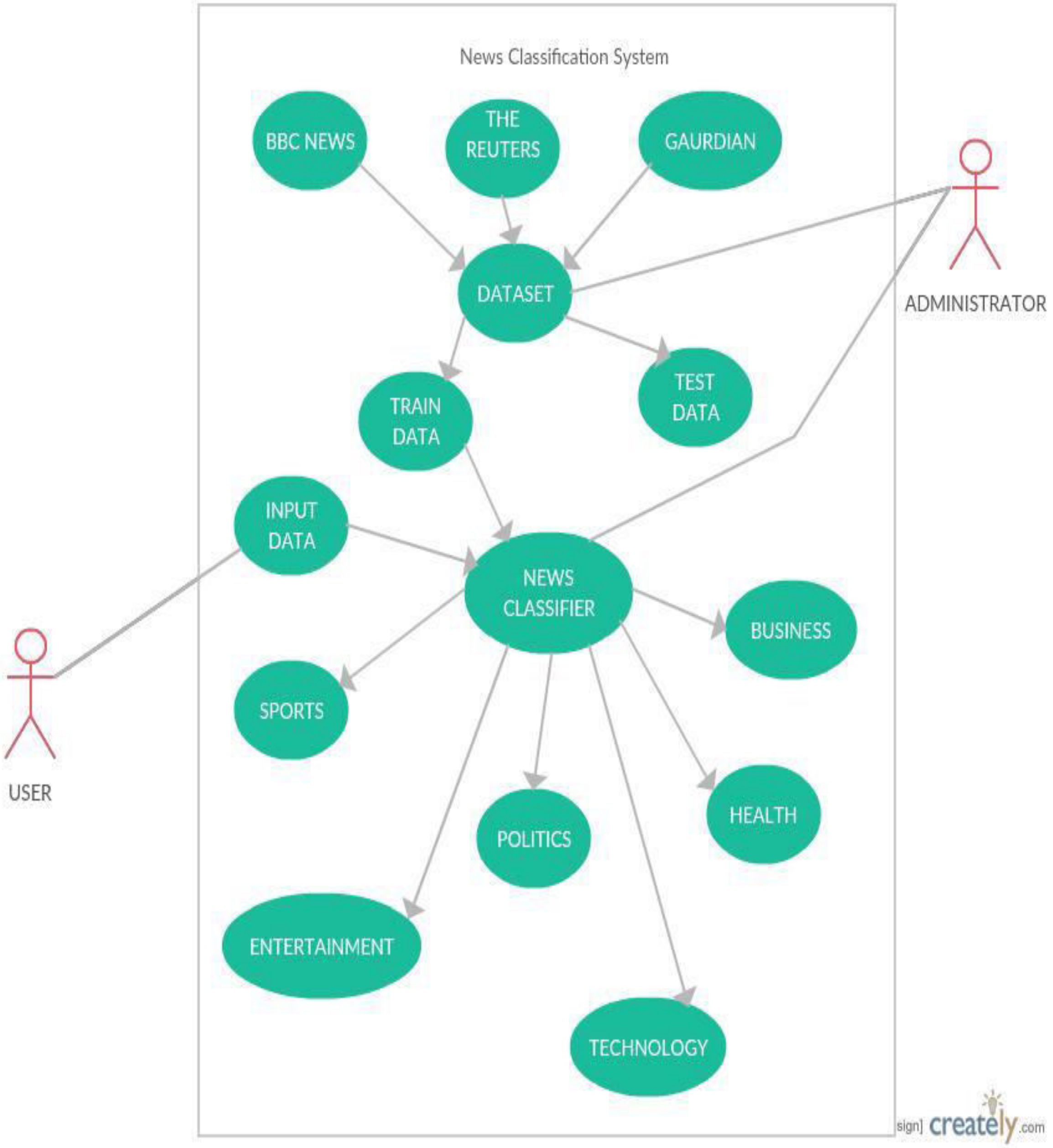


Figure 3: Detailed Use Case Diagram of the project

2.4.2. Class diagram

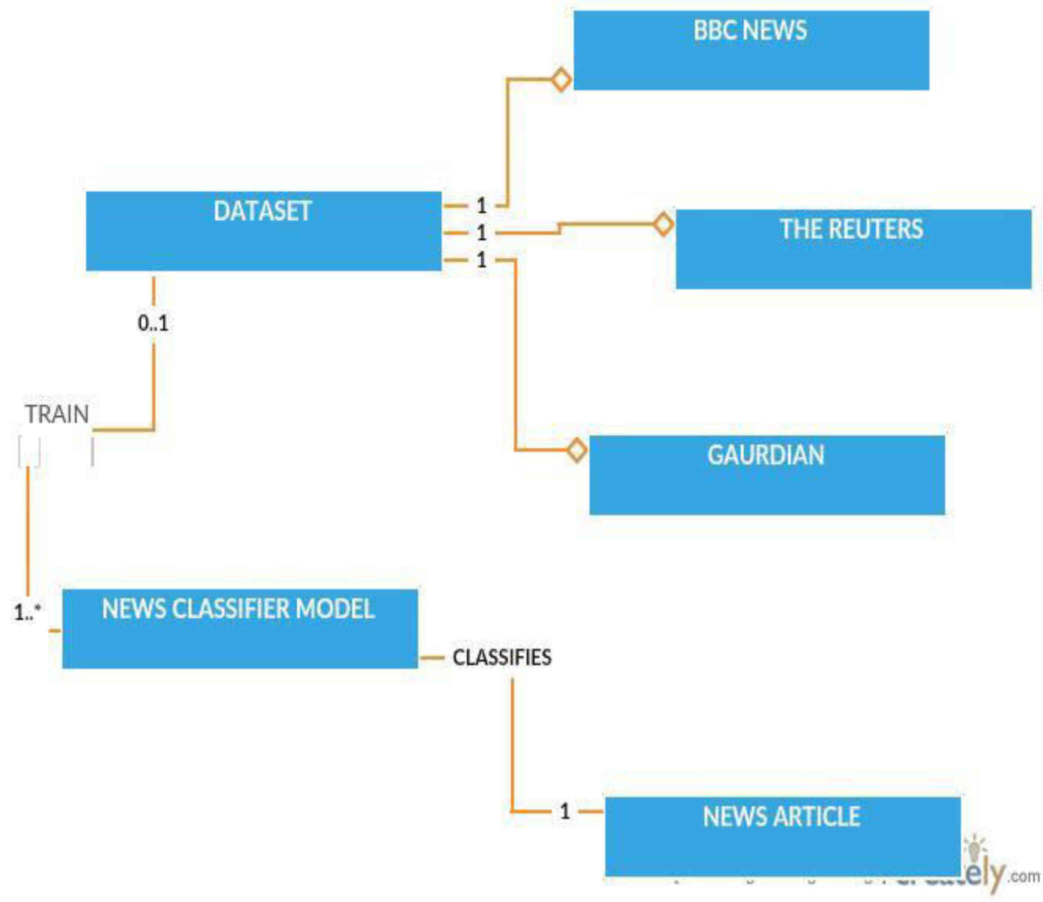


Figure 4- Elucidated Class Diagram

2.4.3. Data Structures and Algorithms

Naive Bayes

The Naive Bayes [7] is a technique that uses the probabilities of each attribute so as to make all variables conditionally independent based on a certain outcome. The Bayesian classifiers primarily utilize supervised learning on training documents to evaluate the variables of the generative model. We start off implementing Naive Bayes by taking the assumption that all the parameters are independent (even though they might actually not be). This assumption forms the basis of this classification and helps to simplify our calculation. Surprisingly, it turns out that in many cases, it actually yields estimates that are on par to the results which we would have acquired from a more (computationally) exorbitant model that acknowledges the conditional dependencies between the parameters.

$$P\left({}^{c}/_{d}\right) = \frac{P\left({}^{d}/_{c}\right)P(c)}{P(d)}$$

- P(c|d) is the posterior probability of class (c) given predictor (d).
- P(c) is the prior probability of class.
- P(d|c) is the likelihood which is the probability of predictor given class.
- P(d) is the prior probability of predictor.

TF-IDF

TF-IDF [8] denotes for Term-Frequency and Inverse Document Frequency. It is a suites of techniques used to calculate similarity between queries and documents. It is quantified by means of the sum of term frequency-like numbers (TFs) multiplied by terms' importance. The term importance is often represented by the IDF (the inverse document frequency). Indeed it is the algorithm of IDF that is used in practice.

Principally, the more recurrent the term is in a document the substantial is the TF coefficient. An opposite scenario holds for the term importance coefficients, which are sizeable for terms that transpire in lesser documents, i.e., more important. Therefore, to determine the value of TF*IDF, it is required to know the number of term occurrences.

$$\boldsymbol{Tfidf(t,d,D) = tf(t,d).idf(t,D)}$$

- Tf = how many times a term can be found in a document or a category
- Idf = log * Inverse Probability of a term being found in an essay

DECISION TREE

A decision tree [8][9] is a structure akin to a flowchart. Each internal node represents a "test" on a particular characteristic such the coin flip and whether the coin flip test will result in heads or tails. Now each branch depicts the result of the test whereas each leaf node represents a class label, that is, decision taken after computing all attributes. Classification rules are illustrated by the path from root to leaf.

Decision analysis and the tools that enable it such as decision tree and influence diagrams are used to evaluate the expected values of various alternatives available.

A common application of decision trees is in the field of operations management. In reality, decisions have to be taken online with insufficient information; hence a decision tree should be coupled with a best choice model such as probability model. Using decision trees as a descriptive means for calculating conditional probabilities is another one of the many uses for decision trees.

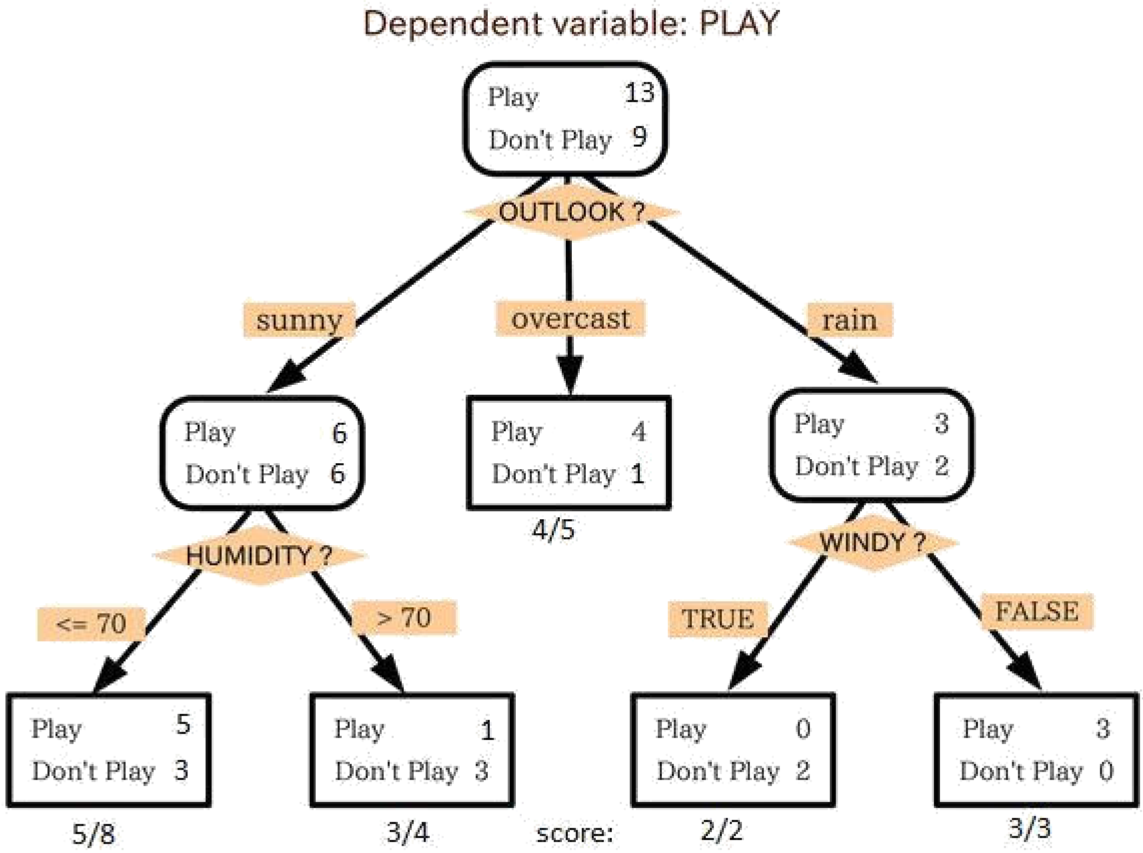


Figure 5- Example of a Decision Tree

K-NN (K – Nearest Neighbour)

For a better understanding of this concept, let us take a simple case to understand this algorithm[9].

Following is a diagram depiction of red circles (RC) and green squares (GS):

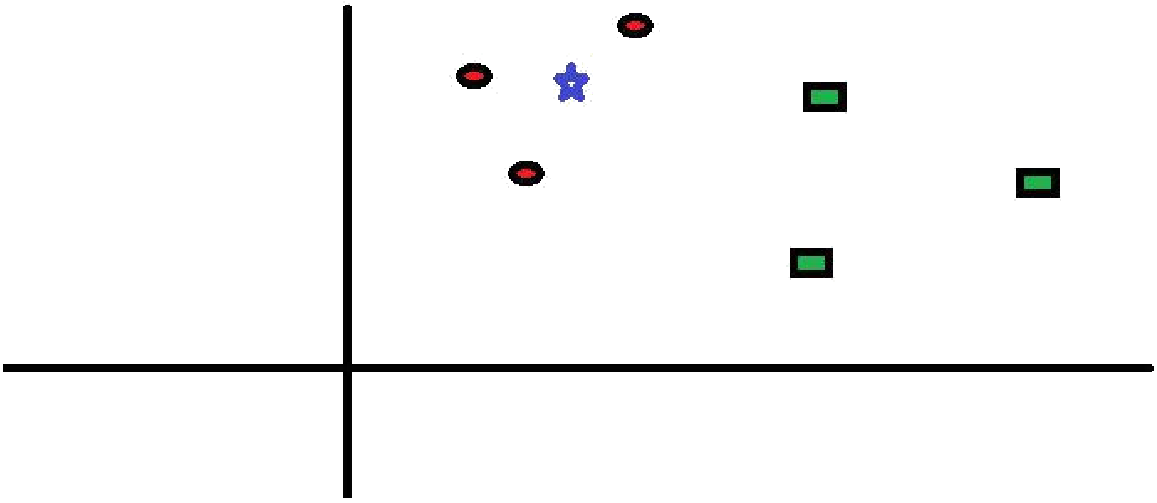


Figure 6- Diagram (i) of Example of K-NN

Our intention is to find out the class of the blue star (BS). Now, BS can either be in Red Circle or Green Squares and nothing else. The "K" in KNN algorithm stands for the nearest neighbours we desire to take vote from. Let's say K = 3. Consequently, we will now make a circle with Blue Star as centre which would be as big as to encompass only three data (as K=3) points on the plane. Kindly refer to the forthcoming diagram for better understanding of the details:

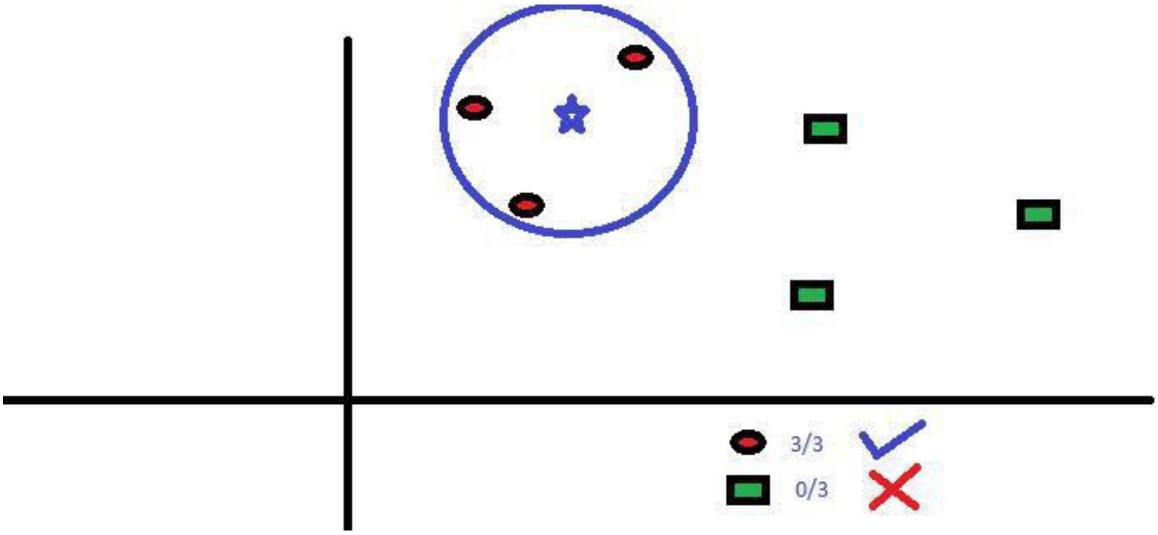


Figure 7- Diagram (ii) of Example of K-NN

If we observe closely, we would realize that the three closest points to Blue Star are all Red Circles. As a result, with good confidence level we can say that the Blue Star should belong to the class Red Circles. In this scenario, the choice became very evident as all three votes from the closest neighbour went to Red Circles. From the above example it becomes pretty noticeable that the choice of the parameter K is very crucial in this algorithm.

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [9], is an unsupervised machine learning technique.It is one of the method used for topic modeling.It considers a document as an amalgam of many subjects and in turn every subject has its own probability distribution of words. LDA allows a set of document to be explicated by unspotted groups and explicate why few divisions of the data arealike. K is the amount of subjects. The probability density is given by-

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1},$$

- **α is an input of the Dirichlet,** that is before**, on the per**-document subjectdistribution
- **θ is subject distribution for document i**

Max Entropy

It is a classifier in probability [10] that is a part of the class of exponential models. For achieving maximum entropy we should have uniform distribution of things or in other words should have the most randomness. Max entropy has been generally utilized for an assortment of common dialect undertakings, including dialect demonstrating, content division, grammatical feature labeling, and prepositional expression connection. The likelihood display for a classifier is restrictive model given by P (C|F1, ..., Fn) over a reliant class variable C speaking to one of the conceivable grouping marks, adapted on a few component variable F1 through Fn.

2.5. Analyzing the risk and Mitigating of the Plan

Risk Id	Risk Describe	Risk Area (Identify Risk Areas for your project)	Probability (P)	Impact (I)	RE (P*I)	Risk for Mitigation (Y/N)	Mitigation Plan if 8 is 'Y'	Contingency plan , if any
1.	Accuracy of query with large data-set	HQL	5	5	25	Y	This can be mitigated by carefully studying the problem and testing the query with different data-set.	
2.	False discoveries are very much possible	Generated Data Sets	3	5	15	Y	This can be mitigated by carefully formulating the query and implementing themodel while testing it rigorously.	

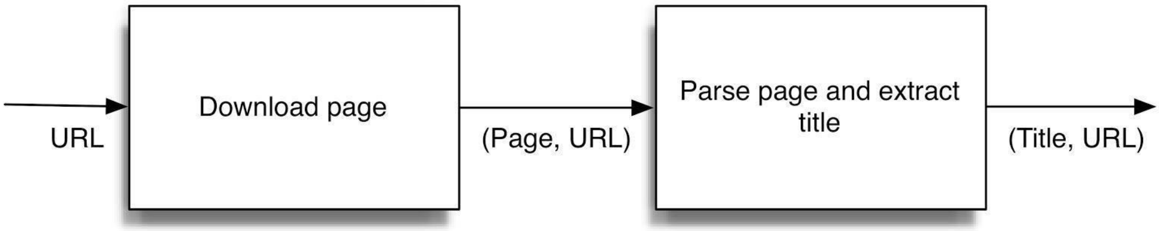
3.	Discriminated Data canlead to improper results	Query	3	4	12	N		Discriminated Data can behandled by proper pre-processing or by totallydiscarding it and usingsome other data set.
4.	Data leak risk	Security	3	5	15	Y	The data- set needs to be properly encrypted and should be carefully handledto prevent any misuse andleaks.	
5.	Hardware Failure	Hardware	1	3	3	N		Careful and regular backups need to be takenwhile performing thecalculations.

Table 1- Risk Analysis

3. Implementation and Testing

3.1. Implementation details and issues

First step was to get the corpus or the dataset which were scrapped using Web Scrapping technique and some of the documents were downloaded from the internet.**Urllib2** library was used to fetch all the URLs and data parsing from these URLs were done through library called **BeautifulSoup**.



We have used the **Natural Language Toolkit** (NLTK) package for text- processing such as **tokenization** and **removal of the stop word**. You assume a document as a string in tokenization and thereafter, divide it into a list of tokens. During removal of the stop words, the documents pertaining to an English stopword list were discarded by us that didn’t add much value to the information like “the”, “to” etc.

For **feature extraction** we have used scikit learn a python library. This revolves around minimizing the resources that we need for explicating a huge data set. We have used Count vectorizer, Tfidf Vectorizer and Hashing Vectorizer for feature extraction. They converted the text document collection to a matrix containing how often the token occurs.

After extracting of the feature, the eminent task of classifying the text is feature selection, whose main intent is choosing a portion of features from the initial document. For this we have used **sklearn.feature_selection** module from scikit learn.

After feature selection next step is to **dividing the documents into already defineddivisions**. The documents are divided as supervised and unsupervised methods. When knowledge of class label relating to document is there, its supervised, otherwise, it is unsupervised classification.

For supervised learning we have trained the Naive bayes classifier ,Decision tree classifier and K-NN classifier using trained data. We have used Scikit learn that contained all the above mentioned techniques classifier like **sklearn.tree. DecisionTreeClassifer, sklearn. NaiveBayesClassifer & sklearn.neighbors. KNeighborsClassifier etc.**. The scikit framework is useful in giving attributes associated with testing accuracy.

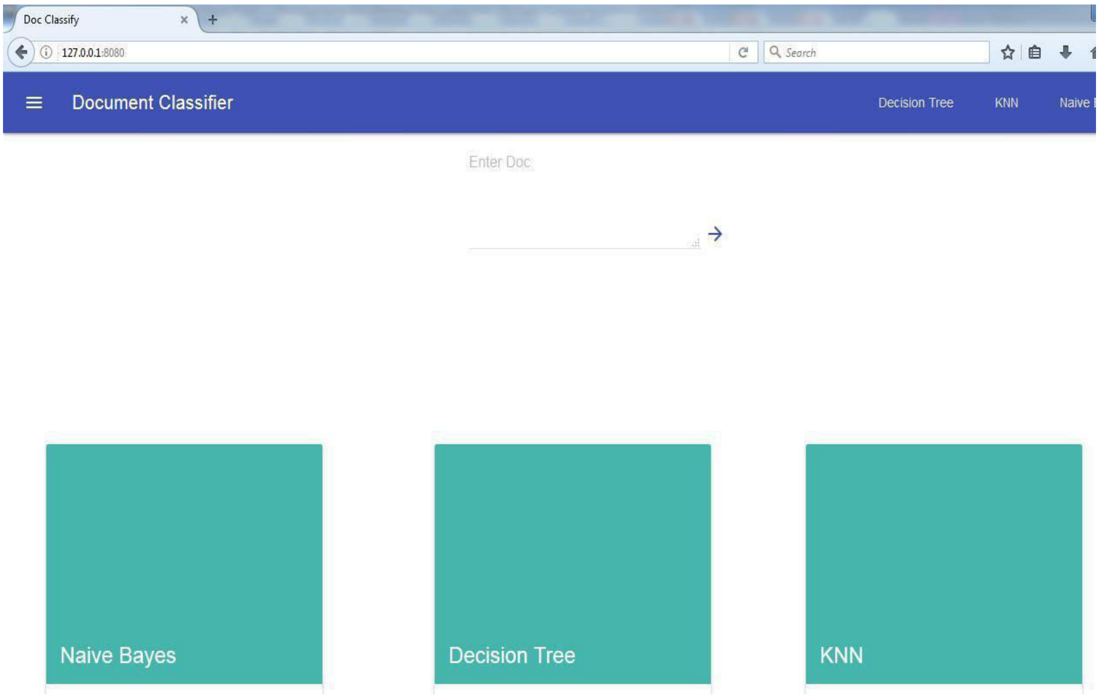
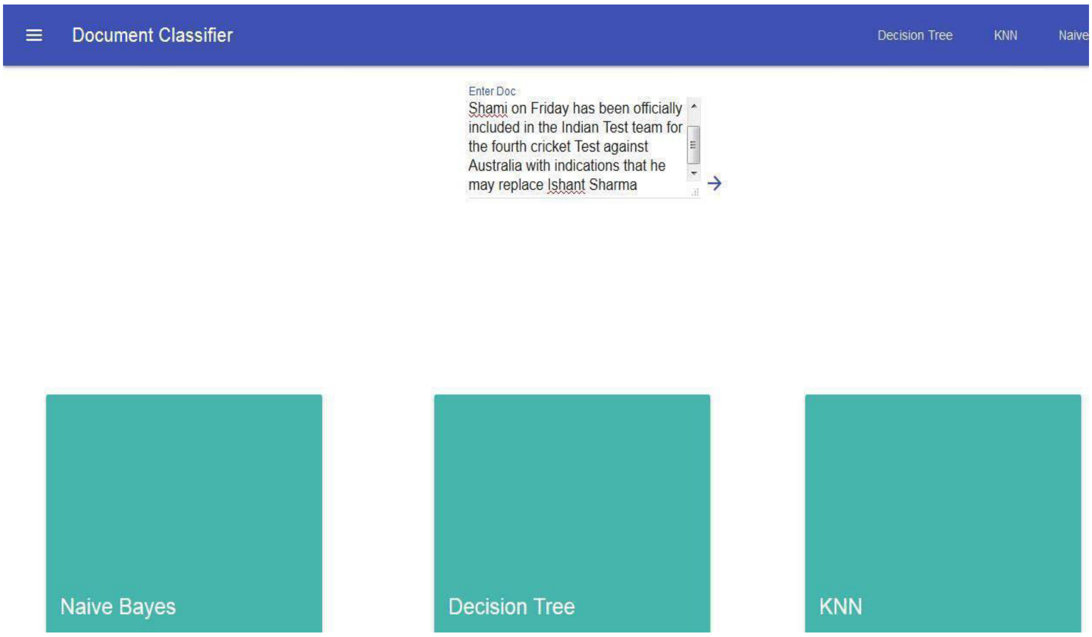


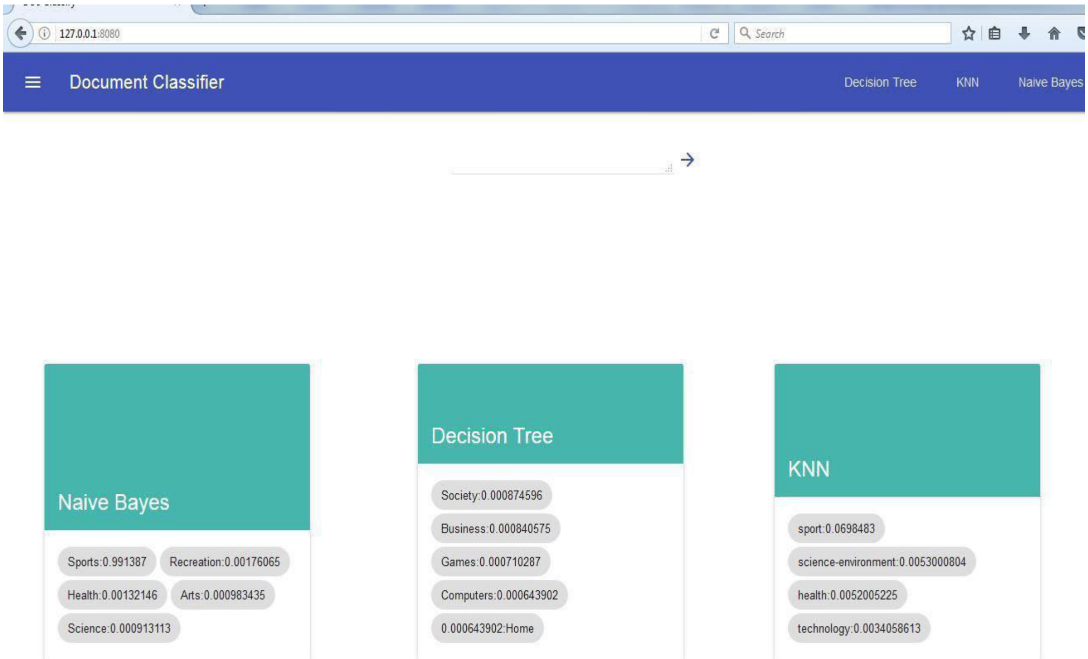
Figure 8 – Project Implementation



Enter Doc

Shami on Friday has been officially included in the Indian Test team for the fourth cricket Test against Australia with indications that he may replace Ishant Sharma

Figure 9 – Project Implementation



Naive Bayes

Sports:0.991387 Recreation:0.00176065
Health:0.00132146 Arts:0.000983435
Science:0.000913113

Decision Tree

Society:0.000874596
Business:0.000840575
Games:0.000710287
Computers:0.000643902
0.000643902:Home

KNN

sport:0.0698483
science-environment:0.0053000804
health:0.0052005225
technology:0.0034058613

Figure 10 – Project Implementation

3.2. Testing

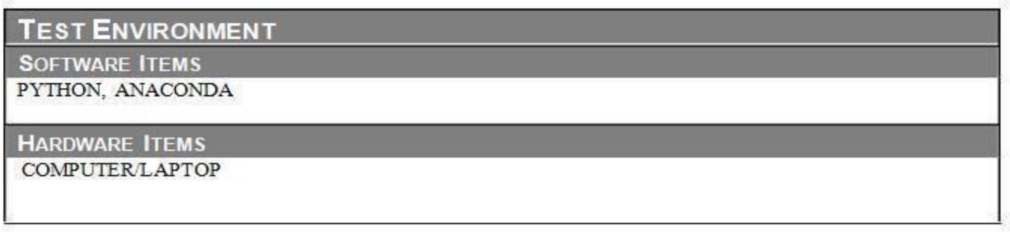
3.2.1. Testing

- The test plan includes testing the project on various data sets and checking if they are accurate as the real life scenario. The test plan also includes white box testing and black box testing. White box testing is a strategy for testing in which the inside structure is checked and code verification is done, remembering the prototype implementation. Here are a few focuses:

Type of Test	**Will Test Be Performed?**	**Comments/Explanations**	**Software Component**
Requirements Testing	Yes	Online articles are tested on our project to determine the accuracy and efficiency	JetBrain-PyCharm
Unit	Yes	Individual categories text files can be tested	JetBrain-PyCharm,Scikit-learn
Integration	Yes	Collection of all files are producing required results	JetBrain-PyCharm,Scikit-learn
Performance	Yes	Our algorithm is 80 to 85 % efficient	JetBrain-PyCharm,Scikit-learn

Stress	No	None	None
Compliance	No	Not Applicable	None
Security	No	Not Applicable	None

Table 2 - Testing Plan



TEST ENVIRONMENT

SOFTWARE ITEMS

PYTHON, ANACONDA

HARDWARE ITEMS

COMPUTER/LAPTOP

Test Schedule			
Activity	**Start Date**	**Completion Date**	**Number of Hours**
Learning Techniques of machine learning	15/12/2016	2/1/2017	216
Used Web Scraping to collect hundreds of articles as sample inputs for the testing	2/1/2017	20/1/2017	200
Used Tf-Idf algorithm on the Sample	20/1/2017	15/2/2017	200
Implemented, Naive Bayes machine learning technique on the input samples	15/2/2107	15/3/2017	250
Implemented Decision tree machine learning technique on input Samples	20/3/2017	17/4/2017	230
Implemented K-NN machine learning technique on input Samples	25/4/2017	2/5/2017	220

Table 3 - Test Schedule

3.2.2. Subdivision of Component and kind of testing needed

S.No	List of the modules needing testing	Kind of Testing needed	Skill to implement testCases
1	Extraction	Requirement	White Box
2	Web Scrapping	Unit	White Box
3	Cleaning dataset	Performance	White Box
4	Naïve Bayes	Performance	Black Box
5	TF-IDF	Integration	White Box
6	Decision Tree	Performance	Black Box
7	K-NN	Performance	Black Box

Table 5- Table containing component decomposition.

3.2.3. Mention the test cases in required format

Test Case id	Test Name	Data to pass as parameter	Desired Output	Result
1.1	Extraction time	Doc file	Less than 1 min	Pass
2.1	Web pages to txt	HTML	Text file	Pass
3.1	Removing stop Words	Text file	Clean data in txt format	Pass
4.1	Classification of article using naïve Bayes	Text	Classified news article	Pass
5.1	Classification of article using Decision Tree	Text	Classified news article	Pass
6.1	Classification of article using K-NN	Text	Classified News Article	Pass

Table 6 - Test cases for each component

4. Findings & Conclusion

4.1. Findings

We have learned new software's like python, anaconda and pycharm. We also found that with the correct component determination and a sufficiently vast training size we can make a classifier to order reports with an exactness of 77% into their individual areas. These outcomes are fascinating both in the practical and theoretical sense. Theoretically, would we be able to immovably make the determination that articles from these distinctive segments are genuinely composed in an unexpected way, and in this manner separate some data that may help journalists composing for a specific area.

4.2. Conclusion

This project was made to analyze the accuracies of various classificators, which in our case are Naïve Bayes, KNN, Decision Tree. I scraped through the web and extracted the relevant documents and kept the number of topics to minimal so as to lessen the power for computing. Then, the major challenge arrived in making the data test ready, where in I had to tune various parameters to obtain the optimal accuracy. Once, the test and train data were ready, using the sci-kit libraries, i was able to obtain the values from all the three models, I tweaked various parameters, mostly the accuracy depended on how well you trained the data. Out of the three ML algorithms, in most number of cases Naïve Bayes predicted the correct results, but KNN and Decision Trees were not far behind.

4.3. Future Work

While different procedures for record classification could be investigated, I accept there is bounty scope for future work with the classification systems. More confounded capabilities could be inspected for Naive Bayes and Maximum Entropy grouping. The capabilities we investigated gave a decent base to grouping however in specific cases, particular components could be added to expand exactness. For instance, the capabilities that work best with daily paper articles specifically can be investigated all the more completely. Additionally work should be possible is hoping to improve preparing on these training set capabilities or astute methods for diminishing the dimensionality of these natural feature sets to be computationally effective and ideally precise.

5. References

[1] Ramdass, Dennis, and Shreyes Seshasai. "Document classification for newspaper articles." (2009).

[2] Kaur, Harmandeep, Sheenam Malhotra, and Fatehgarh Sahib. "Online News Classification: A Review." *International journal of Innovation inEngineering and Technology (IJIET)* 2.2 (2013).

[3] Bai, Yiqi, and Jie Wang. "News classifications with labeled LDA."

Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on. Vol. 1.SCITEPRESS, 2015.

[4] Lam, Wai, and Kei Shiu Ho. "FIDS: an intelligent financial Web news articles digest system." *IEEE Transactions on Systems, Man, andCybernetics-Part A: Systems and Humans* 31.6 (2001).

[5] Hakim, Ari Aulia, et al. "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach." *Information Technology and Electrical Engineering(ICITEE), 2014 6th International Conference on*. IEEE, 2014.

[6] Yang, Yiming, et al. "Learning approaches for detecting and tracking news events." (2000).

[7] Shimodaira, Hiroshi. "Text Classification using Naive Bayes."
Learning and Data Note 7 (2014).

[8] Ramos, Juan. "Using tf-idf to determine word relevance in document queries."
Proceedings of the first instructional conference on machine learning. 2003.

[9] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003):

[10] Nigam, Kamal, John Lafferty, and Andrew McCallum.
"Using Maximum Entropy for Text Classification."

[11] Bird, Steven. "NLTK: the natural language toolkit.

" *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, 2006.