

NEWS CLASSIFICATION

K. Naveen Kumar

ID No: B121250

Murali Chowhan

ID No: B121251

Under the guidance of

Mr. P. Laxmi Narayana

Assistant Professor in the Dept. of CSE



**Department of Computer Science and Engineering
RAJIV GANDHI UNIVERSITY OF KNOWLEDGE
TECHNOLOGIES , BASAR
TELANGANA - 504107**

NEWS CLASSIFICATION

*Project Report submitted to
Rajiv Gandhi University of Knowledge Technologies, Basar
for the partial fulfillment of the requirements
for the award of the degree of*

**Bachelor of Technology
in
Computer Science and Engineering**

by

**Kammari Naveen Kumar (B121250)
Murali Chowhan (B121251)**

Under the guidance of

**Mr. P. Laxmi Narayana
Assistant Professor in the Dept. of CSE**



**Department of Computer Science and Engineering
RAJIV GANDHI UNIVERSITY OF KNOWLEDGE
TECHNOLOGIES, BASAR
APRIL 2018**



**Department of Computer Science and Engineering
RAJIV GANDHI UNIVERSITY OF KNOWLEDGE
TECHNOLOGIES, BASAR**

CERTIFICATE

This is to certify that the Project Report entitled ‘**NEWS CLASSIFICATION**’ submitted by **K. Naveen Kumar(B121250) & Murali Chowhan (B121251)**, Department of Computer Science and Engineering, Rajiv Gandhi University Of Knowledge Technologies, Basar; for partial fulfillment of the requirements for the degree of Bachelor of Technology in Computer Science and Engineering; is a bonafide record of the work and investigations carried out by them under my supervision and guidance.

Project Supervisor
Mr. P. Laxmi Narayana
Assistant Professor

External Examiner

Head of the Department
Mr. G. Ranjith Kumar
Assistant Professor



Department of Computer Science and Engineering
RAJIV GANDHI UNIVERSITY OF KNOWLEDGE
TECHNOLOGIES, BASAR

DECLARATION

We hereby declare that the work which is being presented in this project entitled, " **NEWS CLASSIFICATION** " submitted to **RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES, BASAR** in the partial fulfillment of the requirements for the award of the degree of **BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING**, is an authentic record of our own work carried out under the supervision of **Mr. P. Laxmi Narayana, Assistant Professor in Department of Computer Science And Engineering, RGUKT, Basar**. The matter embodied in this project report has not been submitted by me/us for the award of any other degree.

Place: **K. Naveen Kumar(B121250)**

Date: **Murali Chowhan (B121251)**

ACKNOWLEDGEMENTS

We take this opportunity to express our deep and sincere gratitude to our Supervisor **Mr. P. Laxmi Narayana** for his valuable guidance and for giving us the opportunity to work under him. His constant encouragement, support and generous attitude were a tremendous boost for our work.

We would like to express our sincere gratitude to our HOD **Mr. G. Ranjith Kumar** for encouraging us . We also express our sincere thanks to our Project Coordinator **Mr. Sujoy Sarkar** for providing guidance during the evaluation processes.

Foremost, We would like to express our sincere gratitude to **Sri Dr. A. ASHOK (I.A.S)**, is the Vice Chancellor of the RGUKT-Basar, for his motivation and providing all facilities to do this project.

We thank all our seniors, friends, and members of our project group in materializing this report and also for the lively support and encouragement given from time to time. We have great regard, and we wish to extend our warmest thanks to our classmates who offered constant support during one year span of dissertation work.

K. Naveen Kumar(B121250)

Murali Chowhan (B121251)

ABSTRACT

In recent years, text mining has gained higher popularity. Generally now a days data is available to us through many resources. This data can be in unstructured form and we have a lot of ways to convert unstructured data into structured data. News articles are one of the most important factors that have influence on various sections. We considered the problem of classifying the topic of news (category) articles for which there are multiple relevant sector labels.

Our idea is to propose a system which will perform its task based on text news articles. Our system has several phases like News Collections to collect news from internet resources, Preprocessing phase for preprocessing the text articles to get useful information, Feature extraction to extract useful metrics for classification and finally the Classification phase will classify the news articles into categories.

In this project we have crawled news articles using several python libraries. We used Natural Language ToolKit for preprocessing of news articles. We used several methods for extracting features from preprocessed data ,like CountVectorizer. For the last phase , we have used Naive Bayes classifier. The result obtained are promising , improved classification accuracy by preprocessing techniques.

TABLE OF CONTENTS

CERTIFICATE.....	iii
DECLARATION.....	iv
ACKNOWLEDGEMENT.....	v
ABSTRACT.....	vi
TABLE OF CONTENTS.....	vii
TABLE OF FIGURES.....	ix
LIST OF TABLES.....	x
CHAPTER 1: INTRODUCTION.....	1
1.1 News Classification.....	1
1.2 Components of system.....	1
1.2.1 News Crawling.....	1
1.2.2 Preprocessing.....	2
1.2.3 Feature Extraction	2
1.2.4 Classification	2
1.3 Objective of the Project.....	3
1.4 Organization of Thesis.....	3
CHAPTER 2: LITERATURE OF REVIEW.....	4
2.1 Stock Trend Prediction Using News Sentiment Analysis	
-Kalyani Joshi , Prof. Bharathi H. N. , Prof. Jyothi Rao.....	4
2.2 Preprocessing Techniques for Text Mining - An Overview	
-Dr. S. Vijayarani , Ms. J. Ilamathi and Ms. Nithya.....	4
2.3 News Classification and Its Techniques: A Review	
-Gurmeet Kaur , Karan Bajaj.....	4

CHAPTER 3: PRE-PROCESSING.....	6
3.1 News Collection.....	6
3.2 Tokenization.....	6
3.3 Removal of stop words.....	7
3.4 Word Stemming.....	8
3.5 Lemmatization.....	8
 CHAPTER 4: PROPOSED MODEL.....	 9
4.1 Feature Extraction.....	9
4.1.1 CountVectorizer.....	9
4.1.2 TF-IDF.....	10
4.2 Naive Bayes Classifier.....	10
4.3 Random Forest Classifier.....	12
4.4 Support Vector Machine Classifier.....	13
 CHAPTER 5: RESULTS AND DISCUSSIONS.....	 14
5.1 News Crawling.....	14
5.2 Feature Extraction.....	14
5.3 Naive Bayes Classifier.....	15
5.4 Support Vector Machine Classifier.....	17
5.5 Random Forest Classifier.....	19
5.6 Comparison between classifiers.....	21
 CHAPTER 6: CONCLUSION	 22
REFERENCES.....	23

LIST OF FIGURES

Figure Number	Name of the Figure	Page Number
1.1	System Design	2
4.1	Countvectorizer Example	9
4.2	Basic Bayes Theorem Formula	11
4.3	Multinomial Bayes Theorem Formula	11
4.4	Random Forest-Decision Tree	12
4.5	Support Vector Machine	13
5.1	Crawling news articles by using URLs	14
5.2	Feature Vector Matrix Example	15
5.3	Input test articles	15
5.4	Output of test articles with class labels	16
5.5	Accuracy of Naive Bayes classifier	16
5.6	Sample test articles for SVM	17
5.7	Output of test articles using SVM	18
5.8	Accuracy of SVM	18
5.9	Input test articles for Random Forest	19
5.10	Output of Random Forest classifier	20
5.11	Accuracy of Random Forest Model	21

LIST OF TABLES

Table Number	Name of Table	Page No.
5.6.1	Accuracy Comparison	21

CHAPTER 1

INTRODUCTION

1.1 News Classification

We have a large amount of information being stored in the electronic format. We could interpret and analyse such data and extract facts that help in decision -making. The news information is quickly and easily available from the last decade. Now news is easily accessible via content providers as online news services. In such text data, huge amount of information exists which can be beneficial in several areas. But , classification is quite a challenging field in text mining. It requires preprocessing to convert data into structured information.

With the increase in the number of news it has got difficult for users to access news of his interest which makes it a necessity to categorize news so that it could be easily accessed. Categorization refers to grouping that allows easier navigation among articles. Internet news needs to be divided into categories. This will help users to access the news of their interest in real-time without wasting any time.

1.2 Components of News Classification System

News Classification consists some steps in order to classify the news articles. News Crawling, Preprocessing of news articles, Feature extraction and Classification are main phases of this system. Input to this is a set of news articles. The output of this system is assignment of class labels for the news articles whose category is not known.

1.2.1 News Crawling

The first step of this system is to collect different types of news articles. There are many resources for news articles. This step extracts news articles from internet resource i.e google news. The news articles will be extracted using appropriate url of news articles.

1.2.2 Preprocessing

After crawling the news articles we need to extract useful information for further phase of our system. We preprocess the news articles in order to remove useless information from news articles. NLTK is used for the preprocessing phase. The output of this phase will contain news articles with useful information.

1.2.3 Feature Extraction

In this phase features will be extracted from preprocessed news articles. Feature extraction is the process to extract the most essential characteristics from the data. The most essential data means that's on the basis of that's the characters can be represented. Appropriate features will be extracted in this phase from preprocessed news articles.

1.2.4 Classification

Classification is a process of identifying category of a new observation of data. We will train our classifier on the basis of training set of data containing observations whose category membership is known. The classification is executed on the basis of stored features.

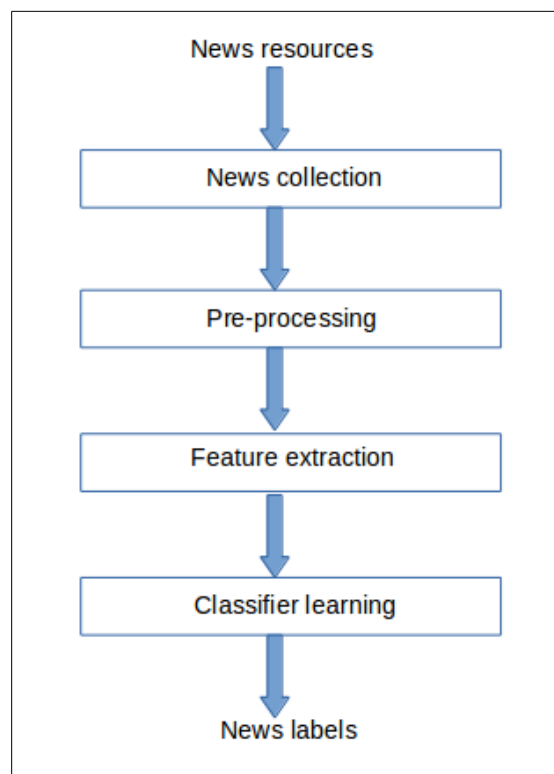


Fig: 1.1 System Design

1.3 Objective of the Project

The objective of the project is to preprocess the news articles and classify them into categories. So that we can easily identify/search a news article in less time. Here our aim is to crawl the news articles from internet and preprocess them to extract useful information and try to classify them based on the features.

1.4 Organization of Thesis

The Remaining Chapters as follows; CHAPTER 2 describes the literature review of the project. CHAPTER 3 describes Preprocessing and its techniques. CHAPTER 4 describes the Proposed model. CHAPTER 5 describes Results and Discussions of the project. CHAPTER 6 describes the conclusion of the project and finally the References.

2.1 Stock Trend Prediction Using News Sentiment Analysis

-Kalyani Joshi , Prof. Bharathi H. N. , Prof. Jyothi Rao

This paper presented an automated text mining based approach to aggregate news stories from various sources and create a News Corpus. The Corpus is filtered down to relevant sentences and analyzed using Natural Language Processing (NLP) techniques. This paper presented an idea to crawl the news articles from internet resources. This paper includes a step by step process to preprocess news articles using preprocessing techniques. We proposed a workflow of our system which comprises phases like news crawling, preprocessing, classifier learning and testing with news data .

2.2 Preprocessing Techniques for Text Mining - An Overview

-Dr. S. Vijayarani , Ms. J. Ilamathi and Ms. Nithya

They had analysed different types of papers and concluded their journal for Text mining, Stemming, Stop words elimination, TF/IDF algorithms, Wordnet, Word Disambiguation. A very good discussion about preprocessing techniques. The paper gives a detailed outline of common stemming techniques and existing stemmers for Indian languages. The purpose of stemming is to reduce different grammatical forms or word forms of a word like its noun, adjective, verb, adverb etc. The goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. This paper discusses different methods of stemming and their comparisons in terms of usage, advantages as well as limitations.

2.3 News Classification and Its Techniques: A Review

-Gurmeet Kaur , Karan Bajaj

This paper discussed that the purpose of classifying text. Collection of news from internet resources , pre-processing of text is discussed. This paper discusses different types of techniques in the part of pre-processing such as stemming, stopword removal,

types of stemming techniques etc. It also discussed about the features extraction from pre processed text . It has explained about different types of features selections like boolean weighting ,Information Gain, Term Frequency Inverse Class Frequency etc. This paper has information about different types of classification algorithms.

Preprocessing is a preliminary processing of data in order to prepare it for the primary processing or for further analysis. It is any type of processing performed on raw data to prepare it for another processing procedure. The preprocessing is an important task and critical step in text mining. Data preprocessing is used for extracting interesting and non-trivial and knowledge from unstructured text data. By using Natural Language Processing (NLP) we will preprocess our text data. The following steps are used for preprocessing of text data.

3.1 News Collection

The first step of news classification is accumulating news from various sources. This data may be available from various sources like newspapers, press, magazines, radio, television and World Wide Web and many more. But we are using google news as resource for crawling news articles. The collection of news article's anchors will be taken from the google news and extracting the news from the urls. We have used python libraries for crawling the news articles from google news. This step includes complexity while dealing with parsing the html pages. After crawling news articles we are saving them into csv file.

3.2 Tokenization

This technique is used to get tokens from set of documents. This step takes the collected news articles as the input and fragment huge text into small tokens. Each word in the news article is treated as a token. The output of this step will be used for upcoming steps in preprocessing. We have used a inbuilt library called NLTK for preprocessing. It provides different techniques for preprocessing of text. It provides word tokenization and sentence tokenization for tokenization technique. We have used word tokenization.

A sample example for this technique will be shown below . Input text is in raw data.

rawdata="This is example for tokenization. Here is a small sentence which will be used as input to tokenization. Word tokenize will split sentence and treats each word as separate token"

After using word_tokenizer from NLTK the output will be contain each word as separate token.

['This', 'is', 'example', 'for', 'tokenization', '.', 'Here', 'is', 'a', 'small', 'sentence', 'which', 'will', 'be', 'used', 'as', 'input', 'to', 'tokenization', '.', 'Word', 'tokenize', 'will', 'split', 'sentence', 'and', 'treats', 'each', 'word', 'as', 'separate', 'token']

3.3 Removal of Stop Words

The stop words are language specific and doesn't carry any information. Stop words doesn't give more effective weight for our classification. Generally, stop words should be removed to prevent them from affecting our results. Stop words can be removed from data in many ways. There removal can be on the basis of concepts i.e. the removal will be of the words which provide very fewer information about classification. Another way of removal of stop words is the removal of the words that are present in the list of English stop words. The list is made up of approximately 545 stop words.

We will collect a set of all stop words from specific language and we take each token from sentence which has completed its tokenization step. We compare each token in the sentence with the set of all stop words. So if the stop words are present in sentence the we will remove the stop words and remaining words will be there in the sentence. Input for this technique is tokenized sentences.

Input sentence to stop word removal technique:

['This', 'is', 'example', 'for', 'tokenization', '.', 'Here', 'is', 'a', 'small', 'sentence', 'which', 'will', 'be', 'used', 'as', 'input', 'to', 'tokenization', '.', 'Word', 'tokenize', 'will', 'split', 'sentence', 'and', 'treats', 'each', 'word', 'as', 'separate', 'token']

The NLTK will provide collection of stop words for english language. After performing this technique the output will be as follows:

['This', 'example', 'tokenization', '.', 'Here', 'small', 'sentence', 'used', 'input', 'tokenization', '.', 'Word', 'tokenizer', 'split', 'sentence', 'treats', 'word', 'separate', 'token']

3.4 Word Stemming

After the removal of stop words the next activity that is performed is stemming . This step reduces a word to its root . The motive behind using stemming is to remove the suffixes so that the number of words would be brought down . For example the words like user, users, used, using all can be reduced to the word 'use' . This will reduce the required time and space . For stemming there exists many stemmers like S-Stemmers , Lovins Stemmer , Porter Stemmer , Porter Stemmer, Paice/Husk Stemmer.

Word stemming is useful to reduce a word into its root word. By having more forms of same word will not give us any useful information. So that the duplication of same word will be eliminated by stemming. Thus, the processing time and the space will be reduced. NLTK will provide several stemmers but PorterStemmer is used due to its accuracy.

Example: Studying,Studied words will be reduced to study (PorterStemmer gives studi as their root word.). Stemming is speed so that we can reduce computational time. But it is not accuracy. Sometimes,it doesn't generate a real word but gives root word.

3.5 Lemmatization

The purpose of Lemmatization is to group together different inflected forms of a word ,called lemma. This process is similar to stemming, as it maps to several words into one common root. The output of lemmatization is a proper word, and basic suffix stripping wouldn't provide the same outcome. Lemmatization is slow when compared with stemming but it is accuracy than the stemming.

In this project we have used sklearn tool to implement the code for all algorithms for feature extraction and to build a classifier.

4.1 Feature Extractions

After preprocessing the news articles, we have to extract features that will be useful in classification. We have used two types of methods to extract features from preprocessed data. CountVectorizer and Tf-Idf are used for feature extraction from preprocessed news articles.

4.1.1 CountVectorizer

The Count Vectorizer provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words, but also to encode new documents using that vocabulary. It convert a collection of text documents to a matrix of token counts.

Algorithm:

1. Extracts all unique words
2. Arrange all unique words as vector
3. Takes each document and encode it as vector ,with length of entire vocabulary and integer count for the number of time a word appeared in the document.
4. Arrange all vectors into the matrix of feature vectors

	are	call	from	hello	home	how	me	money	now	tomorrow	win	you
0	1	0	0	1	0	1	0	0	0	0	0	1
1	0	0	1	0	1	0	0	1	0	0	2	0
2	0	1	0	0	0	0	1	0	1	0	0	0
3	0	1	0	1	0	0	0	0	0	1	0	1

Fig: 4.1 CountVectorizer Example

4.1.2 TF-IDF:

Term Frequency-Inverse Document Frequency (TF-IDF) is a numerical measure that is used to score the importance of a word in a document based on how often it appears in that document and a given collection of documents. The intuition for this measure is, If a word appears frequently in a document then it should be important and we should give that word a high score. But if a word appears in too many other documents, it's probably not a unique identifier, therefore we should assign a lower score to that word. The math formula for this measure is

$$\text{TF-IDF Score} = \text{TF}(x,y) * \text{IDF}$$

$$\text{TF-IDF Score} = \text{TF}(x,y) * \log(N/df)$$

Where, $\text{TF}(x,y)$ = No. of occurrences of x in y document

N = Total number of documents in corpus

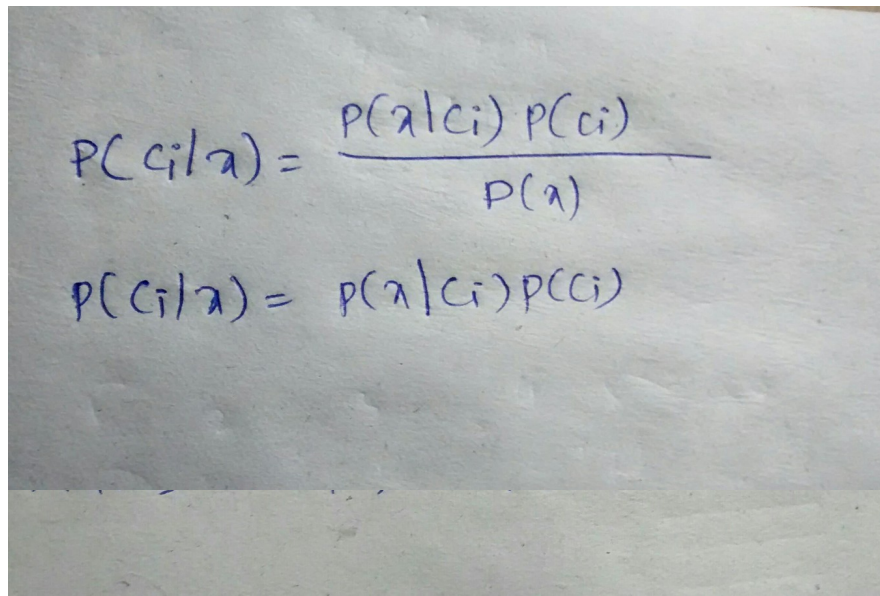
df = No. of documents containing x

4.2 Naive Bayes Classifier

Naive Bayes classifiers, a family of classifiers that are based on the popular Bayes' probability theorem, are known for creating simple yet well performing models, especially in the fields of document classification and disease prediction. In machine learning, Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features. Naive Bayes models are also known under a variety of names in the literature, including simple Bayes and independence Bayes.

In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naïve Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters necessary for classification.

In this project we have used Multinomial Naive Bayes. This classifier attempts to use the frequency of words in a document relative to all the classes which we intend to represent, to determine the class of an unlabeled document. All features are independent of each other. Let's establish the basic classification model. Mathematical formula for this classifier is,



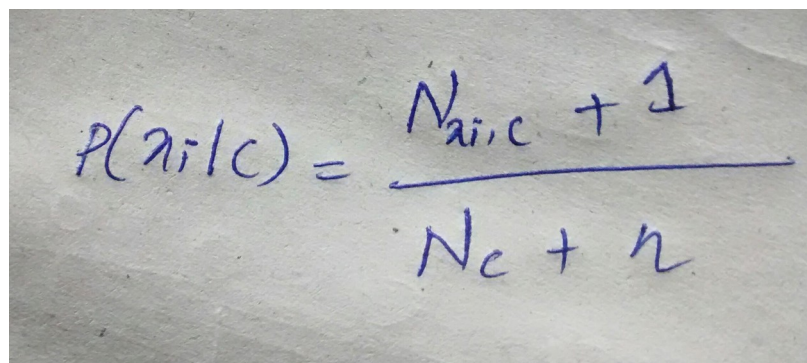
$$P(C_i|x) = \frac{P(x|C_i) P(C_i)}{P(x)}$$

$$P(C_i|x) = P(x|C_i) P(C_i)$$

Fig: 4.2 Basic Bayes Theorem Formula

Where, X= feature vector

Ci= Class labels



$$P(x_i|c) = \frac{N_{x_i,c} + 1}{N_c + n}$$

Fig: 4.3 Multinomial Naive Bayes Formula

Where, $N_{x_i,c}$ = No.of occurrences of x_i word in class c

N_c = No.of words in class c

n = Total No.of unique words in corpus

Multinomial Naive Bayes will assign class label to the unknown document to the class, which has higher probability.

Algorithm:

We need to find the probability of feature vector for all classes using above formula and assign class label to the new article by choosing the class label for which we get higher probability.

4.3 Random Forest classifiers

Random forest is one of machine learning algorithm for classification. This classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. Random forest algorithm is a supervised classification algorithm. As the name suggest , this algorithm creates the forest with a number of trees.

In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy.

Algorithm:

1. Building several decision trees from training dataset
2. Ensemble method for combining trees to produce strong learner
3. Test the new data observation by pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in.

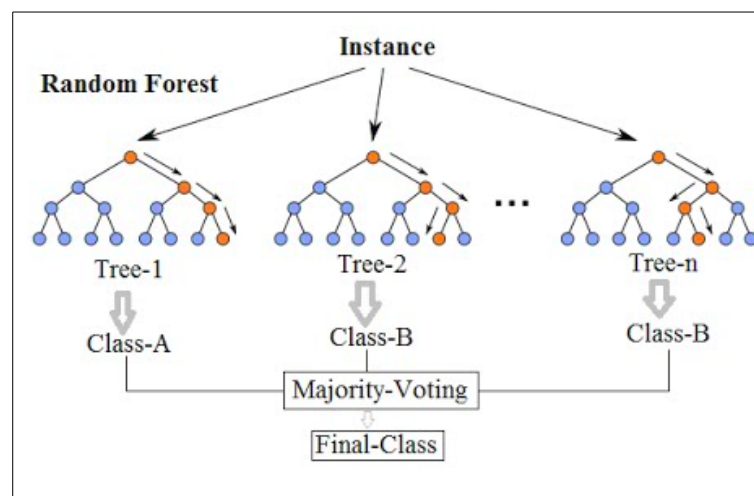


Fig: 4.4 Random Forest-Decision Tree

4.4 Support Vector Machine

A Support Vector Machine is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data, the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side. The SVM is supervised learning.

The SVM classifier is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. Below figure is sample example of two class label data representation in SVM model.

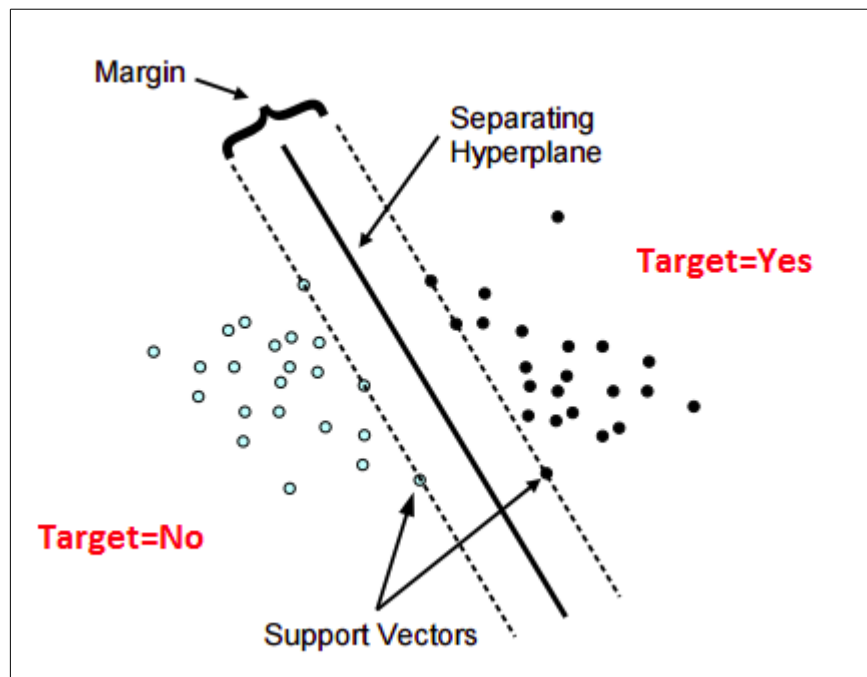


Fig: 4.5 Support Vector Machine

CHAPTER 5

RESULTS AND DISCUSSIONS

We have crawled news articles , preprocessed them and applied proposed models with different test articles. We finally achieve some output for each phase of our project. These are:

5.1 News Crawling

Collecting news articles from google news website , using urls of news pages.

```
https://timesofindia.indiatimes.com/business/india-business/tpg-backed-manipal-acquires-fortis-health-srl-diagnostics/articleshow/63495248.cms
https://timesofindia.indiatimes.com/business/india-business/tpg-backed-manipal-acquires-fortis-health-srl-diagnostics/articleshow/63495248.cms
https://economictimes.indiatimes.com/industry/healthcare/biotech/healthcare/fortis-healthcare-to-demerge-hospital-division-into-manipal-hospitals/articleshow/63493986.cms
http://www.business-standard.com/article/pti-stories/fortis-board-approves-demerger-of-its-hospital-biz-118032800014_1.html
https://www.indiatoday.in/education-today/news/story/indian-railways-over-2-crore-candidates-apply-for-one-lakh-posts-1198674-2018-03-27
https://www.indiatoday.in/education-today/news/story/indian-railways-over-2-crore-candidates-apply-for-one-lakh-posts-1198674-2018-03-27
https://www.thebetterindia.com/135717/govt-offers-10-lakh-prize-for-ideas-improve-indian-railways-services/
http://www.zeebiz.com/india/news-indian-railways-prize-offer-earn-rs-10-lakh-heres-what-you-should-do-check-indianrailwaysgovin-40681
https://scroll.in/latest/873489/mumbai-hdfc-removes-anti-homeless-spikes-outside-its-branch-after-criticism-on-social-media
https://scroll.in/latest/873489/mumbai-hdfc-removes-anti-homeless-spikes-outside-its-branch-after-criticism-on-social-media
https://timesofindia.indiatimes.com/india/following-twitter-outrage-hdfc-bank-removes-metal-spikes-installed-outside-mumbai-branch/articleshow/63477636.cms
https://www.thestatesman.com/cities/mumbai-hdfc-bank-removes-anti-homeless-spikes-after-public-outrage-1502611396.html
http://www.dnaindia.com/business/report-bandhan-bank-lists-with-27-gains-is-bigger-than-bank-of-baroda-pnb-2598435
http://www.dnaindia.com/business/report-bandhan-bank-lists-with-27-gains-is-bigger-than-bank-of-baroda-pnb-2598435
http://business-standard.com/search?type=news&q=bank
http://www.business-standard.com/article/companies/bandhan-bank-outpaces-bob-pnb-on-debut-stock-gains-25-over-issue-price-118032701368_1.html
https://www.livemint.com/Money/80WRH1MeSPuCLU5fh8LubN/Bandhan-Bank-jumps-30-on-stock-market-debut.html
http://economictimes.indiatimes.com/topic/NSE
https://economictimes.indiatimes.com/small-biz/startups/newsbuzz/day-after-selling-se-asia-biz-to-grab-uber-to-intensify-battle-against-ola/articleshow/63475127.cms
https://economictimes.indiatimes.com/small-biz/startups/newsbuzz/day-after-selling-se-asia-biz-to-grab-uber-to-intensify-battle-against-ola/articleshow/63475127.cms
https://www.livemint.com/Companies/IosvBZE42XwSr5cD2iTrL/Why-Ola-is-in-Norwest-Venture-Partners-antiportfolio.html
https://www.entrepreneur.com/article/311111
https://www.wsj.com/articles/why-apple-got-schooled-by-google-1522180531
https://www.wsj.com/articles/why-apple-got-schooled-by-google-1522180531
https://deepmind.com/blog/wavenet-generative-model-raw-audio/
http://www.newindianexpress.com/world/2018/mar/27/senate-committee-summons-facebook-google-twitter-ceos-to-testify-1793252.html
https://www.4hotellers.com/features/article/11140
https://cloud.google.com/text-to-speech/
https://www.livemint.com/Opinion/PXuSxzy9yVvR7NctDQjMMO/The-hidden-pitfalls-of-digital-regulation.html
http://www.financialexpress.com/industry/rs-2500-cr-tax-issue-massive-setback-for-tech-firm-cognizant-as-income-tax-dept-freezes-its-mumbai-chennai-accounts/1113176/
http://www.financialexpress.com/industry/rs-2500-cr-tax-issue-massive-setback-for-tech-firm-cognizant-as-income-tax-dept-freezes-its-mumbai-chennai-accounts/1113176/
http://www.dnaindia.com/business/report-cognizant-s-rs-2500-crore-bank-acs-attached-by-it-department-2598438
```

Fig: 5.1 Crawling news articles by using URLs

5.2 Feature Extraction

We have taken preprocessed text and extracted useful features from it

Sample Input Text

['This is a document of the corpus.','It has been used for the test and train data.','In this document several test words are used and document contains several important content']

Output for input text

```
Feature names
[u'contains', u'content', u'corpus', u'data', u'document', u'important', u'seveal', u'test', u'training', u'used', u'words']

Feature vector matrix
[[0 0 1 0 1 0 0 0 0 0 0]
 [0 0 0 1 0 0 0 1 1 1 0]
 [1 1 0 0 2 1 1 1 0 1 1]]
Total features extracte 11
```

Fig 5.2: Feature Vector Matrix Example

5.3 Naive Bayes Classifier

First we have trained the Classifier with training data and then tested with test data of news articles. It produces accuracy of 85%-94%.

Sample input news articles for testing

```
0 The has booked former chairman-cum-managing di...
1 NEW YORK, April 13 (Reuters) - The U.S. Treasu...
2 The US Treasury added India to its watch list ...
3 NEW DELHI: Cryptocurrency exchange Coinsecure ...
4 The Indian cryptocurrency exchange Coinsecure,...
5 NEW DELHI: In order to bring greater transpare...
6 INTUC said althiugh it was not keen on withdra...
7 Penbrook Management LLC purchased a new positi...
8 News headlines about Ford Motor (NYSE:F) have ...
9 NEW DELHI: Tax authorities have asked field of...
10 Trucks at a toll plaza. Under a proposed natio...
11 NEW DELHI: India has one of the lowest telecom...
12 After debuting successfully in 2017, India is ...
13 By Kate Holton\n\nLONDON (Reuters) - Martin So...
14 Sorrell will be treated as having retired, bas...
15 Martin Sorrell has stepped down as chief execu...
16 firm (ISec) on Saturday reported an exponentia...
17 ICICI Securities (I-Sec), the recently listed ...
18 ICICI Securities's board recommended a final d...
19 The applicable slab rate with respect to an in...
20 The process for filing income tax return (ITR)...
21 Media stories about Goldman Sachs (NYSE:GS) ha...
22 MUMBAI: Tata Motors , which acquired Jaguar La...
23 Shares of Tata Motors Ltd. have fallen 23 perc...
24 As leaks and rumours are majorly putting adver...
25 Apple had launched the new iPhone 8 and 8 Plus...
26 Apple Inc. warned employees to stop leaking in...
27 There is no dearth of spy images of the upcomi...
28 New spyshots of the 2018 Hyundai Santro hatchb...
29 The RTR 160 has been a stepping stone into the...
..
250 luminal\n\nBRCA\n\nWalter and Eliza Hall Insti...
251 Two months after his son, Apartim Dey Singha, ...
252 New York: In a first, biomedical engineers gre...
253 The over 40-yr-old site receives about 250-ton...
254 Shivani Vora, Apr 13 2018, 19:20 IST\n\nTravel...
```

Fig 5.3:Input test articles

Sample output of the test articles

```

0 The has booked former chairman-cum-managing di... 1
1 NEW YORK, April 13 (Reuters) - The U.S. Treasu... 1
2 The US Treasury added India to its watch list ... 1
3 NEW DELHI: Cryptocurrency exchange Coinsecure ... 1
4 The Indian cryptocurrency exchange Coinsecure,... 1
5 NEW DELHI: In order to bring greater transpare... 1
6 INTUC said althiugh it was not keen on withdra... 1
7 Penbrook Management LLC purchased a new positi... 1
8 News headlines about Ford Motor (NYSE:F) have ... 1
9 NEW DELHI: Tax authorities have asked field of... 1
10 Trucks at a toll plaza. Under a proposed natio... 1
11 NEW DELHI: India has one of the lowest telecom... 1
12 After debuting successfully in 2017, India is ... 5
13 By Kate Holton\n\nLONDON (Reuters) - Martin So... 1
14 Sorrell will be treated as having retired, bas... 1
15 Martin Sorrell has stepped down as chief execu... 1
16 firm (ISec) on Saturday reported an exponentia... 1
17 ICICI Securities (I-Sec), the recently listed ... 1
18 ICICI Securities's board recommended a final d... 1
19 The applicable slab rate with respect to an in... 1
20 The process for filing income tax return (ITR)... 1
21 Media stories about Goldman Sachs (NYSE:GS) ha... 1
22 MUMBAI: Tata Motors , which acquired Jaguar La... 1
23 Shares of Tata Motors Ltd. have fallen 23 perc... 1
24 As leaks and rumours are majorly putting adver... 5
25 Apple had launched the new iPhone 8 and 8 Plus... 5
26 Apple Inc. warned employees to stop leaking in... 5
27 There is no dearth of spy images of the upcomi... 5
28 New spyshots of the 2018 Hyundai Santro hatchb... 5
29 The RTR 160 has been a stepping stone into the... 5
.. ... ..
250 luminal\n\nBRCA\n\nWalter and Eliza Hall Insti... 6
251 Two months after his son, Apartim Dey Singha, ... 6
252 New York: In a first, biomedical engineers gre... 6
253 The over 40-yr-old site receives about 250-ton... 6
254 Shivani Vora, Apr 13 2018, 19:20 IST\n\nTravel... 4

```

Fig 5.4: Output of test articles with class labels

```

[280 rows x 2 columns]
[[23  0  0  1  0]
 [ 3 63  0  3  0]
 [ 1  2 63  0  0]
 [ 0  1  0 72  0]
 [ 0  1  1  5 41]]

('Accuracy score: ', 0.9357142857142857)

```

Fig 5.5: Accuracy of Naive Bayes Classifier

5.4 Support Vector Machine Classifier

We have used SVM as our classifier to compare accuracy of it with Naive Bayes Classifier. The result of SVM is mentioned below:

Sample input test articles

```
0 The has booked former chairman-cum-managing di...
1 NEW YORK, April 13 (Reuters) - The U.S. Treasu...
2 The US Treasury added India to its watch list ...
3 NEW DELHI: Cryptocurrency exchange Coinsecure ...
4 The Indian cryptocurrency exchange Coinsecure,...
5 NEW DELHI: In order to bring greater transpare...
6 INTUC said althiugh it was not keen on withdra...
7 Penbrook Management LLC purchased a new positi...
8 News headlines about Ford Motor (NYSE:F) have ...
9 NEW DELHI: Tax authorities have asked field of...
10 Trucks at a toll plaza. Under a proposed natio...
11 NEW DELHI: India has one of the lowest telecom...
12 After debuting successfully in 2017, India is ...
13 By Kate Holton\n\nLONDON (Reuters) - Martin So...
14 Sorrell will be treated as having retired, bas...
15 Martin Sorrell has stepped down as chief execu...
16 firm (ISec) on Saturday reported an exponentia...
17 ICICI Securities (I-Sec), the recently listed ...
18 ICICI Securities's board recommended a final d...
19 The applicable slab rate with respect to an in...
20 The process for filing income tax return (ITR)...
21 Media stories about Goldman Sachs (NYSE:GS) ha...
22 MUMBAI: Tata Motors , which acquired Jaguar La...
23 Shares of Tata Motors Ltd. have fallen 23 perc...
24 As leaks and rumours are majorly putting adver...
25 Apple had launched the new iPhone 8 and 8 Plus...
26 Apple Inc. warned employees to stop leaking in...
27 There is no dearth of spy images of the upcomi...
28 New spyshots of the 2018 Hyundai Santro hatchb...
29 The RTR 160 has been a stepping stone into the...
..
250 luminal\n\nBRCA\n\nWalter and Eliza Hall Insti...
251 Two months after his son, Apartim Dey Singha, ...
252 New York: In a first, biomedical engineers gre...
253 The over 40-yr-old site receives about 250-ton...
254 Shivani Vora, Apr 13 2018, 19:20 IST\n\nTravel...
```

Fig 5.6: Sample test articles for SVM

Sample output of SVM classifier

		news category
0	The has booked former chairman-cum-managing di...	1
1	NEW YORK, April 13 (Reuters) - The U.S. Treasu...	1
2	The US Treasury added India to its watch list ...	1
3	NEW DELHI: Cryptocurrency exchange Coinsecure ...	1
4	The Indian cryptocurrency exchange Coinsecure,...	1
5	NEW DELHI: In order to bring greater transpare...	1
6	INTUC said althiugh it was not keen on withdra...	1
7	Penbrook Management LLC purchased a new positi...	1
8	News headlines about Ford Motor (NYSE:F) have ...	1
9	NEW DELHI: Tax authorities have asked field of...	1
10	Trucks at a toll plaza. Under a proposed natio...	1
11	NEW DELHI: India has one of the lowest telecom...	1
12	After debuting successfully in 2017, India is ...	5
13	By Kate Holton\n\nLONDON (Reuters) - Martin So...	1
14	Sorrell will be treated as having retired, bas...	1
15	Martin Sorrell has stepped down as chief execu...	1
16	firm (ISec) on Saturday reported an exponentia...	1
17	ICICI Securities (I-Sec), the recently listed ...	1
18	ICICI Securities's board recommended a final d...	1
19	The applicable slab rate with respect to an in...	1
20	The process for filing income tax return (ITR)...	1
21	Media stories about Goldman Sachs (NYSE:GS) ha...	1
22	MUMBAI: Tata Motors , which acquired Jaguar La...	1
23	Shares of Tata Motors Ltd. have fallen 23 perc...	1
24	As leaks and rumours are majorly putting adver...	5
25	Apple had launched the new iPhone 8 and 8 Plus...	5
26	Apple Inc. warned employees to stop leaking in...	5
27	There is no dearth of spy images of the upcomi...	5
28	New spyshots of the 2018 Hyundai Santro hatchb...	5
29	The RTR 160 has been a stepping stone into the...	5
..
250	luminal\n\nBRCA\n\nWalter and Eliza Hall Insti...	6
251	Two months after his son, Apartim Dey Singha, ...	6
252	New York: In a first, biomedical engineers gre...	5
253	The over 40-yr-old site receives about 250-ton...	5
254	Shivani Vora, Apr 13 2018, 19:20 IST\n\nTravel...	4

Fig 5.7: Output of test articles using SVM

```
[280 rows x 2 columns]
Confusion Matrix:
[[23  0  0  1  0]
 [ 2 65  0  2  0]
 [ 1  0 64  1  0]
 [ 0  0  0 73  0]
 [ 0  2  2 19 25]]

Accuracy:  0.892857142857
```

Fig 5.8: Accuracy of SVM

5.5 Random Forest Classifier

We have used RF model as our classifier to compare accuracy of it with Naive Bayes Classifier. The result of RF model is mentioned below:

Sample input test articles

```
0 The has booked former chairman-cum-managing di...
1 NEW YORK, April 13 (Reuters) - The U.S. Treasu...
2 The US Treasury added India to its watch list ...
3 NEW DELHI: Cryptocurrency exchange Coinsecure ...
4 The Indian cryptocurrency exchange Coinsecure,...
5 NEW DELHI: In order to bring greater transpare...
6 INTUC said althiugh it was not keen on withdra...
7 Penbrook Management LLC purchased a new positi...
8 News headlines about Ford Motor (NYSE:F) have ...
9 NEW DELHI: Tax authorities have asked field of...
10 Trucks at a toll plaza. Under a proposed natio...
11 NEW DELHI: India has one of the lowest telecom...
12 After debuting successfully in 2017, India is ...
13 By Kate Holton\n\nLONDON (Reuters) - Martin So...
14 Sorrell will be treated as having retired, bas...
15 Martin Sorrell has stepped down as chief execu...
16 firm (ISec) on Saturday reported an exponentia...
17 ICICI Securities (I-Sec), the recently listed ...
18 ICICI Securities's board recommended a final d...
19 The applicable slab rate with respect to an in...
20 The process for filing income tax return (ITR)...
21 Media stories about Goldman Sachs (NYSE:GS) ha...
22 MUMBAI: Tata Motors , which acquired Jaguar La...
23 Shares of Tata Motors Ltd. have fallen 23 perc...
24 As leaks and rumours are majorly putting adver...
25 Apple had launched the new iPhone 8 and 8 Plus...
26 Apple Inc. warned employees to stop leaking in...
27 There is no dearth of spy images of the upcomi...
28 New spyshots of the 2018 Hyundai Santro hatchb...
29 The RTR 160 has been a stepping stone into the...
..
250 luminal\n\nBRCA\n\nWalter and Eliza Hall Insti...
251 Two months after his son, Apartim Dey Singha, ...
252 New York: In a first, biomedical engineers gre...
253 The over 40-yr-old site receives about 250-ton...
254 Shivani Vora, Apr 13 2018, 19:20 IST\n\nTravel...
```

Fig 5.9: Input test articles for RF

Sample output of the test articles

1	NEW YORK, April 13 (Reuters) - The U.S. Treasu...	1
2	The US Treasury added India to its watch list ...	1
3	NEW DELHI: Cryptocurrency exchange Coinsecure ...	1
4	The Indian cryptocurrency exchange Coinsecure,...	1
5	NEW DELHI: In order to bring greater transparen...	1
6	INTUC said althiugh it was not keen on withdra...	1
7	Penbrook Management LLC purchased a new positi...	1
8	News headlines about Ford Motor (NYSE:F) have ...	1
9	NEW DELHI: Tax authorities have asked field of...	1
10	Trucks at a toll plaza. Under a proposed natio...	1
11	NEW DELHI: India has one of the lowest telecom...	5
12	After debuting successfully in 2017, India is ...	5
13	By Kate Holton\n\nLONDON (Reuters) - Martin So...	1
14	Sorrell will be treated as having retired, bas...	1
15	Martin Sorrell has stepped down as chief execu...	1
16	firm (ISec) on Saturday reported an exponentia...	1
17	ICICI Securities (I-Sec), the recently listed ...	1
18	ICICI Securities's board recommended a final d...	1
19	The applicable slab rate with respect to an in...	1
20	The process for filing income tax return (ITR)...	1
21	Media stories about Goldman Sachs (NYSE:GS) ha...	1
22	MUMBAI: Tata Motors , which acquired Jaguar La...	1
23	Shares of Tata Motors Ltd. have fallen 23 perc...	1
24	As leaks and rumours are majorly putting adver...	5
25	Apple had launched the new iPhone 8 and 8 Plus...	5
26	Apple Inc. warned employees to stop leaking in...	5
27	There is no dearth of spy images of the upcomi...	5
28	New spyshots of the 2018 Hyundai Santro hatchb...	5
29	The RTR 160 has been a stepping stone into the...	5
..
250	luminal\n\nBRCA\n\nWalter and Eliza Hall Insti...	6
251	Two months after his son, Apartim Dey Singha, ...	6
252	New York: In a first, biomedical engineers gre...	6
253	The over 40-yr-old site receives about 250-ton...	6
254	Shivani Vora, Apr 13 2018, 19:20 IST\n\nTravel...	5
255	Ah, there's no time like your twenties. It's t...	5
256	For anyone who wants to bring the gym directly...	5
257	more-in\n\nKarthikeyan's muscular arms give aw...	6
258	'Diet' isn't exactly a word that conjures imag...	5
259	Before advocating any one of the two, just thi...	5
260	for too long may increase the risk of in middl...	6

Fig 5.10: Output of the RF Classifier

```

[280 rows x 2 columns]
Confusion Matrix:
[[22  0  0  2  0]
 [ 0 63  0  6  0]
 [ 1  0 65  0  0]
 [ 0  0  3 70  0]
 [ 1  1  0 13 33]]

Accuracy:  0.903571428571

```

Fig 5.11: Accuracy of RF model

5.6 Comparison between Classifiers

No.of Test articles are: 280

No.of Train articles are:2937

Classifier	Accuracy
Naive Bayes Classifier	94%
Support Vector Machine (SVM)	89%
Random Forest	90%

5.6.1 Accuracy Comparison

CHAPTER-6

CONCLUSION

We are classifying news articles by crawling from internet resources ,preprocessing the news articles and extracting the features from text data and classifying them into categories. Every minute, a lot of news will be generated and this will make difficulty for users to find their preferable news articles. It is difficulty to find/search for a desired news in this huge amount of news. Thus it leads us to classify the news articles and make a way for the users to find their news articles easily.

We have proposed Naive Bayes Classifier as model but we have checked accuracy of it with SVM and RF classifiers. We have tested the accuracy of classifiers with preprocessed news articles and without preprocessed news articles. The accuracy is improved with preprocessing of news articles.

REFERENCES

- (1) Feldman, R., & Sanger, J. (n.d.). Text Mining Preprocessing Techniques. The Text Mining Handbook, 57-63. doi:10.1017/cbo9780511546914.004
- (2) Joshi, K., H. N, B., & Rao, J. (2016). Stock Trend Prediction Using News Sentiment Analysis. International Journal of Computer Science and Information Technology, 8(3), 67-76. doi:10.5121/ijcsit.2016.8306
- (3) Natural Language Toolkit — NLTK 3.2.5 documentation. (n.d.). Retrieved from <https://www.nltk.org/>
- (4) scikit-learn: machine learning in Python — scikit-learn 0.16.1 documentation. (n.d.). <http://scikit-learn.org/>
- (5) <http://blog.so8848.com/2015/08/text-preprocessing-with-python.html#.WtZNN3Vuab->
- (6) A Review on Automatic News Classification using the Probabilistic Classification Algorithms. (2016). International Journal of Science and Research (IJSR), 5(6), 1391-1395. doi:10.21275/v5i6.nov164530
- (7) NLP Tutorial Using Python NLTK (Simple Examples) - Like Geeks. (2017, September 21). Retrieved from <https://likegeeks.com/nlp-tutorial-using-python-nltk>
- (8) 4.2. Feature extraction — scikit-learn 0.19.1 documentation. (n.d.). Retrieved April 19, 2018, from http://scikit-learn.org/stable/modules/feature_extraction.html
- (9) Understanding Support Vector Machine algorithm from examples (along with code). (2018, 15). Retrieved from <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>
- (10) Koehrsen, W. (2017, December 27). Random Forest Simple Explanation ? William Koehrsen ? Medium. Retrieved from <https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d>

(11) Rush, L. (n.d.). TF-IDF | 5 Algorithms Every Web Developer Can Use and Understand. Retrieved from <https://lizrush.gitbooks.io/algorithms-for-webdevs-ebook/content/chapters/tf-idf.html>

(12) Machine Learning :: Text feature extraction (tf-idf) – Part I | Terra Incognita. (n.d.). Retrieved from <http://blog.christianperone.com/2011/09/machine-learning-text-feature-extraction-tf-idf-part-i/>

(13) Lan, H. (2017, August 14). Decision Trees and Random Forests for Classification and Regression pt.1. Retrieved from <https://towardsdatascience.com/decision-trees-and-random-forests-for-classification-and-regression-pt-1-dbb65a458df>